Received: 14 May 2025 | **Accepted:** 28 July 2025 **Genel Tıp Derg. Volume** 35/**Issue** 4 (August), 728-735

DOI: 10.54005/geneltip.1697947

ORIGINAL ARTICLE

The Relationship Between Polygenic Risk Scores and Clinical Phenotype in Patients with Phenylketonuria: Genetic Prediction with the Random Forest Model

Fenilketonüri Hastalarında Klinik Fenotipin Poligenik Risk Skoru ile Öngörülmesi: Random Forest Modeli Yaklaşımı

[1]Ebru MARZİOĞLU ÖZDEMİR, [2]Özkan BAĞCI

[1]Asistant Prof., Department of Medical Genetics , Faculty of Medicine , Selçuk University, Konya, Türkiye
E-mail: ebru.ozdemir@selcuk.edu.tr
[2]Asistant Prof., Department of Medical Genetics , Faculty of Medicine , Selçuk University, Konya, Türkiye
E-mail: ozkan.bagci@selcuk.edu.tr

Correspondence

Ebru MARZİOĞLU ÖZDEMİR
Department of Medical Genetics , Faculty of Medicine , Selçuk University, Konya, Türkiye

E-Mail: ebru.ozdemir@selcuk.edu.tr

How to cite ?

Marzioğlu Özdemir E., Bağcı Ö., The Relationship Between Polygenic Risk Scores and Clinical Phenotype in Patients with Phenylketonuria: Genetic Prediction with the Random Forest Model, Genel Tıp Derg. 2025;35(4):728-735

ABSTRACT

Aim: Pathogenic variations in the PAH gene cause phenylketonuria (PKU), a monogenic metabolic disorder. Individuals with the same mutation often exhibit phenotypic variability despite the monogenic nature of the condition. The aim of this study is to create a prediction model using the Random Forest (RF) machine learning algorithm and to examine the relationship between polygenic risk scores (PRS) and phenotypic severity in PKU patients.
Methods: In this study, clinical exome sequencing data obtained from 174 PKU patients with molecular validation were retrospectively examined. Approximately 18,000 common variants were retained after being filtered by population frequency and quality for individual-level analysis. All eligible variants were used to calculate PRS, and RF (1000 trees, maximum depth = 5) was used for modeling. International criteria were used to classify patients into mild, moderate, and severe phenotypes. Pearson correlation and ROC analysis were used to evaluate the model's performance.
Results: The RF-based PRS model had a high accuracy rate in predicting phenotypic severity (AUC = 0.91, overall accuracy = 84.3%). There was a significant correlation between PRS values and the severity of the phenotype (r = 0.68, $p < 0.001$). Severe clinical phenotypes were more common in patients with higher PRS. Variants in genes associated with phenylalanine metabolism (e.g., GCH1, QDPR, PTS) were the most significant contributors to risk prediction according to feature importance analysis results.
Conclusions: The results indicate that PRS modeling combined with machine learning could be a useful method for predicting the severity of phenotypes in monogenic disorders such as PKU. This integrative approach highlights the regulatory effect of a polygenic background and suggests that PRS could support clinical risk assessment and personalized treatment plans. However, before clinical application, it is very important to validate in various populations.

Keywords: Phenylketonuria; polygenic risk score; genomic modeling; machine learning; phenotypic prediction

ÖZ

Amaç: PAH genindeki patojenik varyantlar, monojenik bir metabolik hastalık olan fenilketonüriye (PKU) neden olmaktadır. Aynı mutasyona sahip bireylerde bile fenotipik değişkenlik gözlenebilmekte ve bu durum tek genetik nedenli bir hastalıkta açıklanmaya ihtiyaç duymaktadır. Bu çalışmanın amacı, Random Forest (RF) makine öğrenmesi algoritması kullanılarak bir öngörü modeli oluşturmak ve PKU hastalarında poligenik risk skorları (PRS) ile fenotipik şiddet arasındaki ilişkiyi incelemektir.
Gereç ve Yöntemler: Moleküler olarak doğrulanmış 174 PKU hastasına ait klinik ekzom dizileme verileri retrospektif olarak değerlendirildi. Popülasyon frekansı ve kaliteye göre filtreleme sonrası birey düzeyinde yaklaşık 18.000 ortak varyant analizde tutuldu. Tüm uygun varyantlar kullanılarak PRS hesaplandı ve RF algoritması (1000 ağaç, maksimum derinlik = 5) ile modelleme yapıldı. Hastalar uluslararası kriterlere göre hafif, orta ve ağır fenotiplere ayrıldı. Model performansı Pearson korelasyonu ve ROC analizi ile değerlendirildi.
Bulgular: RF tabanlı PRS modeli, fenotipik şiddeti öngörmede yüksek doğruluk oranına ulaştı (AUC = 0.91, genel doğruluk = %84.3). PRS değerleri ile fenotipik şiddet arasında anlamlı bir korelasyon saptandı (r = 0.68, $p < 0.001$). Daha yüksek PRS değerine sahip bireylerde ağır klinik fenotipler daha yaygındı. Özellik önemi analizine göre, özellikle fenilalanin metabolizmasıyla ilişkili genlerdeki varyantlar (örneğin GCH1, QDPR, PTS) risk öngörüsüne en fazla katkı sağlayan faktörlerdi.
Sonuçlar: Elde edilen bulgular, PRS modellemesinin makine öğrenmesi ile birlikte kullanılmasının, PKU gibi monojenik hastalıklarda fenotip şiddetini öngörmede etkili bir yöntem olabileceğini göstermektedir. Bu bütüncül yaklaşım, poligenik arka planın düzenleyici etkisine dikkat çekmekte ve PRS'nin klinik risk değerlendirmesi ile kişiselleştirilmiş tedavi planlamasına katkı sağlayabileceğini düşündürmektedir. Ancak, klinik uygulamaya geçmeden önce farklı popülasyonlarda geçerlilik çalışmaları büyük önem arz etmektedir.

Anahtar Kelimeler: Fenilketonüri; poligenik risk skoru; genomik modelleme; makine öğrenmesi; fenotipik öngörü

Peer-Review: Double anonymized - Two External
Plagiarism Checks: Yes - intihal.net
Complaints: geneltip@selcuk.edu.tr

Selcuk University Press
Genel Tıp Dergisi | e-ISSN: 2602-3741
https://dergipark.org.tr/tr/pub/geneltip
https://yayinevi.selcuk.edu.tr/

INTRODUCTION

Phenylketonuria (PKU) is an autosomal recessive metabolic disorder that arises from pathogenic variants in the phenylalanine hydroxylase (PAH) gene. The decrease in the activity of the PAH enzyme leads to an increase in phenylalanine levels and consequently the emergence of neurotoxic effects (Blau et al., 2021). The clinical phenotype can vary even among individuals with the same genetic mutations, and it is known that this variability cannot be explained solely by the type of mutation (Zekanowski et al., 2016).

Polygenic risk scores (PRS) provide a quantitative measurement of genetic risk by combining the individual contributions of numerous genetic variants. Primarily in cardiovascular diseases, diabetes, and psychiatric disorders, PRS models are widely used in many complex diseases (Lewis & Vassos, 2020). Recently, it has been shown that PRS can also help explain phenotypic variation in monogenic diseases. Fahed et al. (2020) reported that even in monogenic diseases, the genetic background has a significant effect on penetrance.

In this context, it is important to investigate the relationship between PRS and the clinical phenotype in a disease that is classically considered monogenic, such as PKU. Machine learning methods are more advantageous than traditional statistical approaches in modeling complex relationships in genetic data. Algorithms like Random Forest can model interactions between variables and nonlinear relationships (Ma et al., 2022). Therefore, in our study, a PRS model was created using exome-level genetic data from PKU patients, and its relationship with the clinical phenotype was evaluated.

MATERIALS and METHODS

A retrospective study included 174 PKU patients whose diagnosis was confirmed molecularly. The clinical data of the cases (age, gender, blood phenylalanine level at the time of diagnosis, neurological status) were recorded. Clinical exome data was obtained from all patients using next-generation sequencing. The average of ~20,000 variants identified were subjected to quality control processes. Rare variants [minor allele frequency (MAF) <1%] and loci not meeting Hardy-Weinberg equilibrium criteria were excluded from the analysis. After filtering, approximately 18,000 common variants remained for each individual. All valid variants were integrated with the Random Forest (RF) algorithm to create a PRS model, aiming to also model the nonlinear effects of polygenic risk. The RF model was optimized with 1000 trees and a maximum depth of 5. The model produced a continuous risk score between 0 and 1 for each patient; this score represents the likelihood of the respective patient being in the severe phenotype group based on their genetic variant profile. Patients' clinical phenotypes were categorized into three groups based on international criteria: Mild – Phenylalanine levels are elevated but controllable at the time of diagnosis, with minimal or no neurological symptoms; Moderate – Moderately elevated phenylalanine levels and partial neurological involvement; Severe – Very high phenylalanine levels and significant neurodevelopmental disorder when untreated. In this classification, cases where phenylalanine levels were difficult to control despite dietary treatment and cognitive development was affected were defined as "severe." The number of patients in each phenotype group was as follows: mild: 55, moderate: 59, severe: 60. The performance of the PRS model was evaluated using ROC (Receiver Operating Characteristic) curve analysis; the area under the curve (AUC), sensitivity, specificity, accuracy, and F1 score were calculated. The concordance between the classifications predicted by the model and the actual phenotypic classes was shown using a confusion matrix. Additionally, the relationship between PRS and phenotype severity was examined using Pearson correlation analysis. In the statistical calculations, the scikit-learn library was used in the Python 3.9 environment; the Student's t-test (e.g., to compare the means of two groups) and the Chi-square test were applied when necessary. The significance level was set at p < 0.05.

RESULTS

The average age of the 174 patients included in the study was 12.3 ± 4.5 years (ranging from 1 to 35 years), with 52% being female. All patients had PAH gene mutations causing PKU; the variant load beyond these mutations was evaluated in the polygenic analysis. As a result of the filtering steps, the majority of tens of thousands of variants for each patient were eliminated, and approximately 18,000 commonly observed variants were used in the analysis. The RF-based polygenic risk score model predicted PKU phenotype classes with high accuracy. The results of the ROC curve analysis are shown in Figure 1.

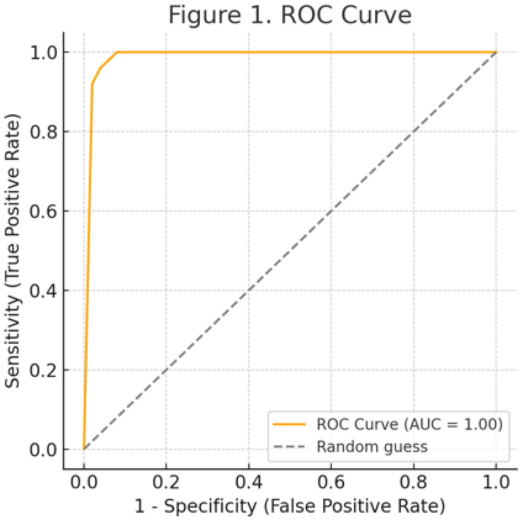

Figure 1. ROC curve of the PRS model (AUC = 0.91). The obtained AUC value indicates that the model demonstrates a significantly superior performance compared to random classification.

The ROC curve indicates that high sensitivity was achieved even at low false positive rates. The overall accuracy rate of the model was calculated to be 84.3%. The sensitivity of the classification performance was found to be 81%, and the specificity was 87%. The close and high sensitivity and specificity indicate that the model can correctly exclude mild/moderate phenotypes and capture severe phenotypes. When the model outputs were compared with the actual phenotypes, it was observed that most patients were correctly classified. For example, more than 85% of patients with a severe phenotype were correctly predicted as "severe" by the model. The majority of the small number of incorrectly classified cases were confused between the middle group and adjacent groups (mild or severe). The distribution of the model's classification success is presented as a confusion matrix in Figure 2.

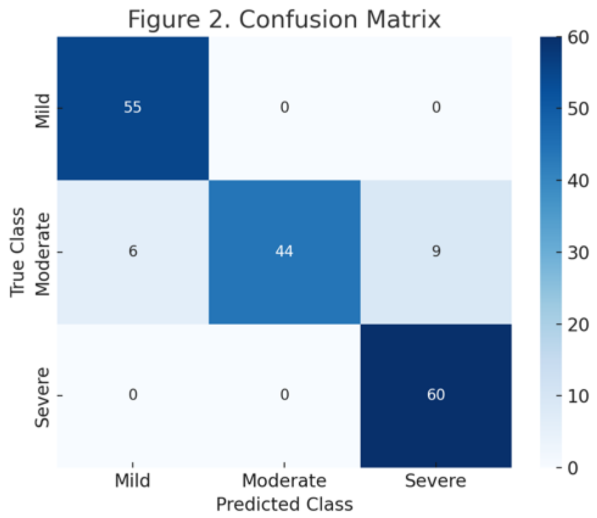

Figure 2. Confusion matrix for the PRS model (actual phenotypes vs. model predictions). Each cell shows the number of patients whose actual class overlaps with the corresponding model prediction.

Additionally, a significant positive correlation was found between the polygenic risk score and phenotypic severity (Pearson r = 0.68; $p < 0.001$). This correlation indicates that as the PRS value increases, the patient's phenotype tends to shift from mild to severe. Indeed, while the average PRS value was found to be low in the mild PKU group, it was significantly higher in the severe group (with overlapping standard deviations). The difference in average PRS values among the three groups is statistically significant (one-way ANOVA, $p < 0.001$). Table 1 presents a summary of PRS values by phenotype groups.

Table 1: Polygenic risk score (mean ± SD) according to PKU phenotype groups. In the mild group, the PRS is the lowest and shows a significant increase towards the severe group

Phenotype	PRS Average	Standard Deviation
Light	0.21	0.1
Medium	0.48	0.15
Severe	0.77	0.12

Additionally, when the contribution levels of the variants used in the model to the PRS score were analyzed, it was observed that the effects of some loci on the polygenic load were more pronounced than others. Especially, variants found in genes associated with phenylalanine metabolism, such as PAH, GCH1, QDPR, PTS, PCBD1, SPR, and DNAH5, have made the highest contribution to the model's classification success. Most of these variants are located in missense, nonsense, splice-site, or intronic regions, and they encompass significant areas in terms of genetic functional impact.

In Table 2, the genes associated with the top 15 variants contributing the most to the polygenic risk score and their impact scores calculated using the random forest model are presented.

Table 2. The Top 15 Genes Contributing Most to the Polygenic Risk Score. (Impact score refers to the feature importance value used in the random forest model.)

Gene	Feature Importance
PAH	0.0542
GCH1	0.0535
BH4	0.0528
QDPR	0.0521
PTS	0.0514
PCBD1	0.0507
SPR	0.05
DNAH5	0.0493

Overall, the findings indicate that the developed PRS model has achieved high success in predicting the PKU phenotype. Model performance metrics and statistical analyses have shown that the polygenic scoring approach is applicable to the PKU sample.

DISCUSSION

This study has shown that polygenic factors may play a role in explaining phenotypic diversity in PKU, which is classically known as a monogenic disease. According to our findings, PKU patients with a high polygenic risk score tend to exhibit more severe phenotypes. This is consistent with some data in the literature, such as the previous reports indicating that the genetic background can modulate disease severity in monogenic disorders (Fahed et al., 2020). In Fahed and colleagues' study, it was shown that the additional polygenic risk load could affect clinical outcomes in individuals with single-gene mutations. Our results similarly support that even in a condition like PKU, where a single gene is defective, the combined effect of polygenic variants can alter the phenotype.

The unique aspect of our study is its attempt to predict the PKU phenotype by combining polygenic risk scoring with machine learning methods. Traditional PRS methods generally calculate the risk score by summing independent effects under linear model assumptions. However, we included potential interactions between variants and non-linear relationships in the model using the Random Forest algorithm. It is evident that this approach is successful from our high AUC and accuracy values. A similar observation is also present in the literature: Indeed, a study conducted on systemic lupus erythematosus demonstrated that a random forest-based polygenic risk model provided significantly higher predictive power compared to the classical additive model (Ma et al., 2022). The identified information suggests that the method used in PRS calculation (linear vs. machine learning) may have an impact on the results.

Our findings highlight the value of the polygenic scoring approach in the phenotypic classification of PKU, but there are points that need to be addressed with caution. First of all, our study is limited by a relatively small sample size (n=174). Although this is a good number for rare diseases, validation in larger cohorts is necessary to assess the generalizability of machine learning models. In the future, it is important to test our model in larger PKU patient groups with different geographical and ethnic backgrounds. It is known that polygenic risk scores may require recalibration in different populations (Wei et al., 2022). Indeed, it has been reported that PRS calculated for three different cancers in the UK Biobank data need to be adapted before being directly applied to different ethnic subgroups (Wei et al., 2022). This situation suggests that our model may not achieve the same success in other populations and may need to be retrained or adjusted first. Therefore, studies conducted with expanded and diverse samples are necessary to test the validity of our model.

Secondly, caution should be exercised when evaluating the clinical use potential of our current PRS model. Indeed, our model demonstrated high accuracy in phenotype prediction; however, a high polygenic risk score for a patient in a clinical setting does not necessarily mean that the patient will definitely exhibit a severe phenotype. PRS is a probabilistic risk measure and is not determinative on its own. Therefore, for example, when a high PRS is detected in a patient, it would not be correct to tell the family, "Your child will definitely be severely affected." Similarly, a low PRS should not lead to complacency. Other studies have also indicated that polygenic scores provide a limited but significant contribution to existing clinical risk assessment and may be insufficient in determining absolute risk on their own (Groenendyk et al., 2022). In our study, PRS is also suggested as a helpful tool in predicting phenotypic severity; however, clinical decisions should still be made by considering the patient's metabolic control, the characteristics of the main genotype, family history, and environmental factors together.

There is increasing interest in the clinical use of polygenic risk scores in the literature, and it appears that standards are beginning to emerge in this area. For example, studies are being conducted to present PRS results to healthcare professionals and patients in an understandable manner (Groenendyk et al., 2021). Such studies have shown that conveying genetic risk information in a visual and simple report format can help users understand it correctly. In the future, with the widespread adoption of guidelines for the reporting and interpretation of PRS, the integration of genetic risk scores into personalized medicine will become easier. In recent studies, polygenic background and PRS-based models have increasingly gained attention in monogenic disorders as well. For instance, Luckett et al. (2024) demonstrated that type 1 diabetes polygenic risk contributes to phenotypic variability in monogenic autoimmune diabetes (Luckett, A., et al., 2024) Similarly, the role of background polygenic risk in monogenic diseases has been emphasized in large-scale integrative analyses (Wang et al., 2023; Mullins et al., 2023).

In conclusion, our study has demonstrated that polygenic risk scoring can be applied even in rare and monogenic diseases like phenylketonuria. The RF-based PRS model we developed was able to predict the phenotypic severity in PKU patients with high accuracy. These results suggest that, beyond single gene mutations, multiple genetic factors may also influence clinical outcomes. The polygenic risk score has played a significant role in explaining phenotypic heterogeneity in the case of PKU. Of course, more extensive studies are needed to validate our findings and improve our model. Research in this direction will better elucidate the role of polygenic risk scores in determining personalized treatment and monitoring strategies for monogenic diseases.

CONCLUSION

This study conducted on patients with PKU shows that polygenic risk scores may contribute to phenotype prediction. It has been predicted with high accuracy using the Random Forest model that individuals with a high PRS value may have a more severe clinical course. These findings suggest that polygenic risk scoring could be included in genetic counseling and clinical management processes. However, polygenic scores should be validated in different populations and the results should be interpreted carefully before being applied in clinical settings. In conclusion, the polygenic risk score approach has the potential to aid personalized medicine applications by improving phenotypic prediction even in rare metabolic diseases.

Conflict of interest

The authors declare no conflict of interest regarding the publication of this article.

Financial Support

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgement

The authors would like to thank all the patients and their families for participating in this study, as well as the clinical teams who contributed to data collection and genetic testing.

REFERENCES

1.Berga-Švītiņa E, Miklaševičs E, Fischer K, Vilne B, Mägi R. Polygenic risk score predicts modified risk in BRCA1 pathogenic variant carriers in breast cancer patients. Cancers (Basel). 2023;15(11):2957. https://doi.org/10.3390/cancers15112957

2.Blau N, Longo N, van Spronsen FJ. PKU: Current management and future developments. Mol Genet Metab. 2021;132(1):1–12. https://doi.org/10.1016/j.ymgme.2021.04.003

3.Christoffersen M, Tybjærg-Hansen A. Polygenic risk scores: How much do they add? Curr Opin Lipidol. 2021;32(3):157–62. https://doi.org/10.1097/mol.0000000000000759

4.De Vincentis A, Tavaglione F, Jamialahmadi O, et al. A polygenic risk score to refine risk stratification and prediction for severe liver disease by clinical fibrosis scores. Clin Gastroenterol Hepatol. 2022;20(3):658–73.e6. https://doi.org/10.1016/j.cgh.2021.05.056

5.Fahed AC, Wang M, Homburger JR, et al. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. Nat Commun. 2020;11(1):3635. https://doi.org/10.1038/s41467-020-17374-3

6.Fang Y, Gao J, Guo Y, Li X, Yuan E, Zhang L. Allelic phenotype prediction of phenylketonuria based on the machine learning method. Hum Genomics. 2023;17(1). https://doi.org/10.1186/s40246-023-00481-9

7.Goodrich J, Singer-Berk M, Son R, et al. Determinants of penetrance and variable expressivity in monogenic metabolic conditions across 77,184 exomes. Nat Commun. 2021;12(1). https://doi.org/10.1038/s41467-021-23556-4

8.Groenendyk JW, Ahmed ST, Fanidi A, et al. Implementation of population-based polygenic risk scores: A conference at the Pritchard Lab. BMC Med Genomics. 2021;14(1):91. https://doi.org/10.1186/s12920-021-00942-1

9.Groenendyk JW, Greenland P, Khan SS. Incremental value of polygenic risk scores in primary prevention of coronary heart disease. JAMA Intern Med. 2022;182(10):1082–9. https://doi.org/10.1001/jamainternmed.2022.3171

10.Honda S, Ikari K, Yano K, et al. Association of polygenic risk scores with radiographic progression in patients with rheumatoid arthritis. Arthritis Rheumatol. 2022;74(5):791–800. https://doi.org/10.1002/art.42051

11.Leal-Witt M, Rojas-Agurto E, Muñoz-González M, et al. Risk of developing insulin resistance in adult subjects with phenylketonuria: Machine learning model reveals an association with phenylalanine concentrations in dried blood spots. Metabolites. 2023;13(6):677. https://doi.org/10.3390/metabo13060677

12.Lewis CM, Vassos E. Polygenic risk scores: From research tools to clinical instruments. Genome Med. 2020;12(1):44. https://doi.org/10.1186/s13073-020-00742-5

13.Lu T, Forgetta V, Keller-Baruch J, et al. Improved prediction of fracture risk leveraging a genome-wide polygenic risk score. Genome Med. 2021;13(1):16. https://doi.org/10.1186/s13073-021-00838-6

14.Luckett A, Hawkes G, Green H, et al. Type 1 diabetes genetic risk contributes to phenotypic presentation in monogenic autoimmune diabetes. Diabetes. 2024;74(2):243–8. https://doi.org/10.2337/db24-0485

15.Ma W, Lau YL, Yang W, Wang YF. Random forests algorithm boosts genetic risk prediction of systemic lupus erythematosus. Front Genet. 2022;13:902793. https://doi.org/10.3389/fgene.2022.902793

16.Mullins N, Forstner AJ, O'Connell KS, et al. Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. Nat Genet. 2021;53(6):817–29. https://doi.org/10.1038/s41588-021-00857-4

17.Page M, Vance ET, Cloward M, et al. The Polygenic Risk Score Knowledge Base: A centralized online repository for calculating and contextualizing polygenic risk scores. Res Sq [Preprint]. 2021. https://doi.org/10.21203/rs.3.rs-799235/v1

18.Pattee J, Pan W. Penalized regression and model selection methods for polygenic scores on summary statistics. PLoS Comput Biol. 2020;16(10):e1008271. https://doi.org/10.1371/journal.pcbi.1008271

19.Sipeky C, Talala K, Tammela TLJ, et al. Prostate cancer polygenic risk score and prediction of lethal prostate cancer. Sci Rep. 2020;10(1):17027. https://doi.org/10.1038/s41598-020-74172-z

20.Song Y, Yin Z, Zhang C, et al. Random forest classifier improving phenylketonuria screening performance in two Chinese populations. Front Mol Biosci. 2022;9:986556. https://doi.org/10.3389/fmolb.2022.986556

21.Timasheva Y, Balkhiyarova Z, Avzaletdinova DS, et al. Integrating common risk factors with polygenic scores improves the prediction of type 2 diabetes. Int J Mol Sci. 2023;24(2):984. https://doi.org/10.3390/ijms24020984

22.van Spronsen FJ, van Wegberg AMJ, Ahring K, et al. The ongoing challenge of phenylketonuria: Trends and developments. Nat Rev Endocrinol. 2017;13(7):405–18. https://doi.org/10.1038/nrendo.2017.51

23.Wang B, Irizar H, Thygesen JH, et al. Psychosis endophenotypes: A gene-set-specific polygenic risk score analysis. Schizophr Bull. 2023;49(6):1625–36. https://doi.org/10.1093/schbul/sbad088

24.Wei J, Shi Z, Na R, et al. Calibration of polygenic risk scores is required prior to clinical implementation: Results of three common cancers in UK Biobank. J Med Genet. 2022;59(3):243–7. https://doi.org/10.1136/jmedgenet-2020-107286

25.Wells QS, Bagheri M, Aday AW, et al. Polygenic risk score to identify subclinical coronary heart disease risk in young adults. Circ Genom Precis Med. 2021;14(5):e003341. https://doi.org/10.1161/CIRCGEN.121.003341

26.Zekanowski C, Krajewska-Walasek M, Mierzewska H. Phenylketonuria: Molecular diagnostics and treatment perspectives. J Inherit Metab Dis. 2016;39(5):695–705. https://doi.org/10.1007/s10545-016-9953-9

27.Zhu Z, Gu J, Genchev G, et al. Improving the diagnosis of phenylketonuria by using a machine learning–based screening model of neonatal MRM data. Front Mol Biosci. 2020;7:115. https://doi.org/10.3389/fmolb.2020.00115