Front. Med. 2025, 19(6): 1060−1069
https://doi.org/10.1007/s11684-025-1183-x

RESEARCH ARTICLE

Risk prediction and therapeutic targets for incident pulmonary hypertension: a large-scale proteomic profiling and Mendelian randomization study

Yuanyuan Zhang[1], Yan Zhang[2], Sisi Yang[1], Yu Huang[1], Yanjun Zhang[1], Ziliang Ye[1], Hao Xiang[1], Xiaoqin Gan[1], Fan Fan Hou (✉)[1], Xianhui Qin (✉)[1]

[1]Division of Nephrology, Nanfang Hospital, Southern Medical University; National Clinical Research Center for Kidney Disease; State Key Laboratory of Multi-Organ Injury Prevention and Treatment; Guangdong Provincial Institute of Nephrology; Guangdong Provincial Key Laboratory of Renal Failure Research, Guangzhou 510515, China; [2]Department of Cardiology, Peking University First Hospital, Beijing 100034, China

Abstract **We aimed to identify plasma proteins associated with pulmonary hypertension (PH) risk, discover potential therapeutic targets for PH, and develop and validate a protein-based prediction model. The development cohort included 38 499 UK Biobank participants from England (split into 70% training and 30% testing set), while the validation cohort comprised 5021 participants from Scotland and Wales. LASSO regression was used to identify predictive proteins in the training set, with model performance assessed using Harrell's C-index, net reclassification improvement (NRI), and integrated discrimination improvement (IDI) in the testing and validation cohorts. We developed a 30-protein risk score, identifying RGMA and NPC2 as causal factors and potential therapeutic targets. Endothelin-1 emerged as a central hub in the protein-protein interaction network. In the testing set, the PH protein risk score demonstrated superior predictive performance for PH risk (C-index = 0.873, 95% CI 0.846–0.900) compared to a basic model (age and sex; C-index = 0.761, 95% CI 0.726–0.795) and a clinical risk model (C-index = 0.843, 95% CI 0.815–0.870). Adding the PH protein risk score to clinical risk factors significantly improved 10-year PH risk reclassification (NRI = 0.258, IDI = 0.053). Similar performance was observed in the validation cohort. These findings underscore the clinical utility of protein biomarkers for PH risk assessment and identify RGMA and NPC2 as promising therapeutic targets.**

Keywords proteomics; incident PH; risk prediction; therapeutic targets; protein–protein interaction network

Introduction

Pulmonary hypertension (PH) represents a progressive and life-threatening cardiopulmonary syndrome characterized by pulmonary vascular remodeling, elevated pulmonary arterial pressure, and increased vascular resistance. These pathological changes culminate in right ventricular dysfunction and ultimately lead to premature mortality [1,2]. The clinical challenge is compounded by the disease's insidious onset and rapid progression, with approximately 50% of cases being diagnosed at advanced stages when therapeutic options are limited [2,3]. This diagnostic delay underscores the critical need for early risk stratification to enable timely clinical intervention and improve patient prognosis.

Current risk prediction models for incident PH predominantly target high-risk populations, such as patients with connective tissue diseases, including systemic sclerosis [4] and systemic lupus erythematosus [5]. Although useful for these specific groups, these models have key limitations: (1) limited generalizability to the general population, where early PH detection remains challenging; (2) suboptimal predictive accuracy due to dependence on routine clinical variables; and (3) a lack of mechanistic insights for therapeutic target discovery. To overcome these constraints, proteomic profiling offers a transformative approach. As functional mediators of biological processes, proteins encapsulate genetic, environmental, and pathological influences [6–8], enabling large-scale plasma proteomic analyses to

Received April 15, 2025; accepted August 27, 2025
Correspondence: Xianhui Qin, pharmaqin@126.com;
Fan Fan Hou, ffhouguangzhou@163.com

both identify novel biomarkers for enhanced risk stratification and uncover molecular mechanisms driving PH pathogenesis.

A recent large-scale proteomic study [9] analyzing plasma samples from 41 931 UK Biobank participants identified 20 potential protein biomarkers for PH prediction. While this represents an important advance in the field, several critical limitations constrain its clinical applicability and biological significance: first, the predictive performance of these proteomic markers was not systematically compared with established clinical risk factors; second, the findings lacked validation in an independent cohort; third, no pathway analyses were conducted to elucidate the biological mechanisms underlying the identified signatures; and fourth, causal inference approaches (e.g., Mendelian randomization (MR) analysis) were not employed to prioritize potential therapeutic targets. These methodological gaps highlight the need for a more comprehensive proteomic investigation that not only identifies predictive biomarkers but also advances our understanding of PH pathogenesis and facilitates target discovery.

The UK Biobank (UKB) Pharma Proteomics Project (UKB-PPP) [10], a sub-study of the UK Biobank, provides an unprecedented opportunity to address these questions through large-scale proteomic analysis. This study sought to identify plasma protein biomarkers associated with PH risk, evaluate their causal relationships and therapeutic potential, and develop and validate a protein-based predictive model while comparing its performance with conventional clinical risk factors in the general population.

Materials and methods

Study design and participants

The UKB is a large, ongoing prospective cohort study initiated between 2006 and 2010 across 22 assessment centers in England, Wales, and Scotland. Approximately 500 000 participants were recruited, who completed touchscreen questionnaires, face-to-face nurse interviews, physical measurements, and provided biological samples for laboratory analyses at baseline [11–13]. The study was approved by the North West Multi-Center Research Ethics Committee (11/NW/0382), and all participants were informed at the start of the study and signed an informed consent.

The UKB-PPP is a substudy within the UKB that conducted plasma proteomic profiling in a subset of over 50 000 participants [10]. From the initial 53 029 participants with complete proteomic data, we excluded 9370 individuals who were of non-White British ancestry, lacked genetic data, showed discrepancies between self-reported sex and X-chromosome heterozygosity, or exhibited excess relatedness. We further excluded 139 participants with prevalent PH at baseline, yielding a final analytical cohort of 43 520 eligible individuals. For analytical purposes, we assigned 38 499 English participants to the model development cohort and 5021 Scottish/Welsh participants to the external validation cohort. Within the development cohort, we performed stratified random sampling based on outcome events to create a training set (70%, n = 26 951) and a testing set (30%, n = 11 548) (Fig. S1).

Blood proteomics assessments

Plasma proteomic profiling was performed using the Olink Explore 3072 proximity extension assay (PEA), an antibody-based platform that quantified 2941 protein analytes, capturing 2923 unique proteins. Prior to analysis, the UKB laboratory team randomized and plated all samples to minimize batch effects. Protein measurements were conducted on three NovaSeq 6000 Sequencing Systems, followed by strict quality control and normalization at Olink's processing facilities. The resulting data were transformed into inverse-rank normalized protein expression (NPX) values for each participant, reported in Olink's proprietary $\log_2$-scale units for relative protein quantification.

Protein selection and risk score derivation

We initially excluded 12 proteins with > 20% missing values, retaining 2,911 proteins for PH risk score development. For these remaining proteins, we imputed the limited missing values using protein-specific mean values, a well-established approach in large-scale proteomic studies [14–16]. A Cox proportional hazards model for PH risk was utilized, incorporating protein levels along with age, and sex as covariates. Variable selection was performed using least absolute shrinkage and selection operator (LASSO) regression with 10-fold cross-validation to optimize penalization strength and determine the final set of predictors (Fig. S2, Table S1). The coefficients for the protein risk score were derived by applying the cross-validated LASSO penalty to the full training dataset within the Cox model's objective function [17]. We then calculated the protein risk score for PH as the weighted sum of the proteins selected by the LASSO regression using the corresponding coefficients as weights.

In addition to the primary holdout method, we performed 5-fold cross-validation to validate the reproducibility of protein biomarker selection (Table S2).

Assessment of clinical risk factors for PH

Clinical risk factors for PH included established

demographic, anthropometric, and clinical factors: age, sex, body mass index (BMI), smoking status, pulse rate, and the prevalence of hypertension, diabetes, cardiovascular disease (CVD), chronic kidney disease (CKD), and chronic respiratory disease.

Data on age, sex, height, weight, and smoking status were collected via standardized questionnaires, while pulse rate was measured during automated blood pressure assessments. BMI was calculated as weight (kg) divided by height squared (m^2). Hypertension was defined as systolic blood pressure (SBP) ⩾ 140 mmHg, diastolic blood pressure (DBP) ⩾ 90 mmHg, self-reported use of antihypertensive, a history of hypertension, or International Classification of Diseases (ICD)-9 (401) or ICD-10 (I10) codes. Diabetes was defined as prevalent diabetes [18] or hemoglobin A1c (HbA1c) ⩾ 6.5%, while CKD included self-reported diagnosis, ICD-10 (N18) codes, estimated glomerular filtration rate (eGFR) < 60 mL/min/1.73 m^2, or urine albumin-to-creatinine ratio (UACR) ⩾ 30 mg/g. eGFR was calculated using the Chronic Kidney Disease Epidemiology Collaboration Equation [19]. CVD comprised coronary heart disease, atrial fibrillation, heart failure, and stroke. Chronic respiratory diseases included chronic obstructive pulmonary disease (COPD), idiopathic pulmonary fibrosis (IPF), and asthma. These conditions were ascertained through self-reported history, hospital admission records, and death registry data.

Study outcome

The study outcome was incident PH, encompassing both primary and secondary PH. Cases were identified through hospital inpatient records and death registry data using ICD-10 codes I27.0 (primary PH) and I27.2 (secondary PH), along with ICD-9 code 4160 (primary PH).

Statistical analysis

The normality of all continuous variables was assessed using the Kolmogorov–Smirnov tests. Non-normally distributed variables (defined as $P < 0.05$; including age, BMI, and pulse rate) were expressed as median (interquartile range, IQR), while categorical variables were presented as proportions. Between-group comparisons (training set vs. testing set in the development cohort, and validation cohort vs. development cohort) were conducted using the Wilcoxon rank-sum tests for continuous variables and the chi-square tests for categorical variables, respectively.

Hazard ratios (HRs) with corresponding 95% confidence intervals (CIs) were calculated using Cox proportional hazards models to examine the association between the PH protein risk score and incident PH risk. Model performance was assessed through both discrimination and calibration measures. Discrimination was evaluated using Harrell's C-index, while reclassification performance and improvement over the reference model were quantified using the continuous net reclassification index (NRI) and integrated discrimination improvement (IDI), computed with the R package survIDINRI. All 95% CIs were estimated using bootstrap methods.

Enrichment analysis was performed on LASSO-selected candidate proteins to investigate potential biological mechanisms. Gene Ontology (GO) functional enrichment and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses were conducted using the Database for Annotation, Visualization and Integrated Discovery (DAVID) database (v2023q4), with statistical significance assessed via Fisher's exact test. Additionally, protein-protein interaction (PPI) networks were constructed and analyzed using the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database.

MR analyses were performed using the "TwoSampleMR" R package to investigate causal relationships between candidate proteins and PH. Protein quantitative trait loci (pQTLs) were identified from the UKB-PPP data [10]. cis-pQTLs were defined as genome-wide significant ($P < 5 \times 10^{-8}$) and linkage disequilibrium (LD)-independent genetic variants located within 500 kb upstream or downstream of the transcription start site of the gene encoding the protein. Genetic information for PH (GWAS ID: finn-b-I9_HYPTENSPUL) was obtained from the IEU GWAS Database (Table S3). MR effects were estimated using the Wald ratio method for single pQTLs and the inverse-variance weighted (IVW) method for multiple pQTLs. The druggability potential of candidate PH-associated proteins was evaluated using the Therapeutic Target Database (TTD; accessed on 21 July 2024).

A two-tailed P-value < 0.05 was considered statistically significant for all analyses. All statistical analyses were conducted using R software (v4.1.1).

Results

Baseline characteristics of participants

Among the 38 499 participants in the development cohort, the mean age was 57.3 years (SD = 8.1), and 46.4% were male. Participants in the training and testing sets exhibited similar baseline characteristics. Compared with the development cohort, participants in the external validation cohort ($N = 5021$) were significantly younger (mean age: 58.0 years vs. 59.0 years, $P < 0.001$), had a higher BMI (27.0 kg/m^2 vs. 26.8 kg/m^2, $P < 0.001$) and a faster pulse rate (69.7 bpm vs. 69.0 bpm, $P < 0.001$). Additionally, the validation cohort exhibited a higher

prevalence of hypertension (60.0% vs. 56.7%, $P < 0.001$) and diabetes (6.4% vs. 6.0%, $P < 0.001$), but a lower prevalence of CKD (7.8% vs. 9.2%, $P < 0.001$) (Table 1).

Association of PH protein risk score with incident PH risk

A PH protein risk score was derived from 30 selected proteins among 2911 candidates in the training set (Fig. S2, Table S1).

In the testing set, over a median follow-up of 13.1 years, 142 (1.2%) participants developed PH. As shown in Fig. 1A, the PH protein risk score was significantly positively associated with incident PH (per SD increment, adjusted HR 2.34, 95% CI 2.05–2.67). Similar results were observed in the external validation cohort (per SD increment, adjusted HR 2.38, 95% CI 1.94–2.92) (Fig. 1B).

Network analysis and biological pathways of PH-associated proteins

Enrichment analysis of the 30 proteins in the PH risk score revealed significant pathway associations across multiple biological domains: extracellular region/space (GO cellular component), cholesterol efflux/blood pressure regulation (GO biological process), hormone activity/calcium ion binding (GO molecular function), and the hypoxia-inducible factor 1 (HIF-1) signaling pathway and vascular smooth muscle contraction (KEGG pathways) (Fig. 2A–2D, Tables S4). PPI analysis identified endothelin-1 (EDN1) as a central network hub (Fig. 2E, Table S5).

MR analysis and druggability of PH-associated proteins

In the two-sample MR analysis, 217 single nucleotide polymorphisms (SNPs) for 28 candidate proteins served as instrumental variables. The results identified significant causal associations: repulsive guidance molecule A (RGMA) showed a positive association with PH (OR 2.73, 95% CI 1.15–6.48), whereas NPC intracellular cholesterol transporter 2 (NPC2) showed an inverse association with PH (OR 0.27, 95% CI 0.08–0.93) (Table S6).

Among the 30-protein risk score panel, nine proteins (116 drugs total) were identified as druggable targets, including: four clinically validated targets (angiopoietin-2 (ANGPT2), carbonic anhydrase 4 (CA4), carbonic anhydrase 14 (CA14), and epidermal growth factor receptor (EGFR)) used in macular degeneration and cancer therapies; three clinical trial targets (RGMA for multiple sclerosis, growth/differentiation factor 15 (GDF15) for heart failure, pro-adrenomedullin (ADM) for respiratory distress syndrome); and two literature-reported targets (NPPB, EDN1) (Table S7). Notably, one drug candidate targeting these proteins was discontinued during Phase 2 trials.

Predictive performance of individual proteins in the PH risk score

Among the 30 candidate proteins evaluated in the testing set, eight demonstrated strong predictive performance for PH risk (C-index ⩾ 0.800), including ANGPT2, GDF15, N-terminal pro-B-type natriuretic peptide (NT-proBNP),

Table 1 Baseline characteristics in development and validation cohorts

Characteristics	Total population			Development cohort		
	External validation cohort ($N = 5021$)	Development cohort ($N = 38\,499$)	*P*-value	Training set ($N = 26\,951$)	Testing set ($N = 11\,548$)	*P*-value
Age, year, median (IQR)	58.0 (51.0, 63.0)	59.0 (51.0, 64.0)	< 0.001	59.0 (51.0, 64.0)	59.0 (51.0, 64.0)	0.724
Male, *n* (%)	2289 (45.6)	17 849 (46.4)	0.301	12 471 (46.3)	5378 (46.6)	0.591
BMI, kg/m^2, median (IQR)	27.0 (24.4, 30.2)	26.8 (24.2, 29.9)	< 0.001	26.8 (24.2, 29.9)	26.7 (24.1, 29.8)	0.774
Pulse rate, bpm, median (IQR)	69.7 (66.0, 73.0)	69.0 (62.0, 76.0)	< 0.001	69.0 (62.0, 76.0)	69.0 (62.0, 76.0)	0.920
Smoking status, *n* (%)			< 0.001			0.954
Never	2804 (55.8)	20 648 (53.6)		14 461 (53.7)	6187 (53.6)	
Former	1612 (32.1)	13 925 (36.2)		9731 (36.1)	4194 (36.3)	
Current	586 (11.7)	3800 (9.9)		2669 (9.9)	1131 (9.8)	
Hypertension, *n* (%)	3013 (60.0)	21 845 (56.7)	< 0.001	15 288 (56.7)	6557 (56.8)	0.995
Diabetes, *n* (%)	323 (6.4)	2303 (6.0)	< 0.001	1604 (6.0)	699 (6.1)	0.928
CVD, *n* (%)	477 (9.5)	3507 (9.1)	0.548	2452 (9.1)	1055 (9.1)	0.380
Chronic respiratory disease, *n* (%)	686 (13.7)	5249 (13.6)	0.957	3688 (13.7)	1561 (13.5)	0.080
CKD, *n* (%)	393 (7.8)	3546 (9.2)	< 0.001	2491 (9.2)	1055 (9.1)	0.865

BMI, body mass index; IQR, interquartile range; PH, pulmonary hypertension; CVD, cardiovascular disease; CKD, chronic kidney disease.

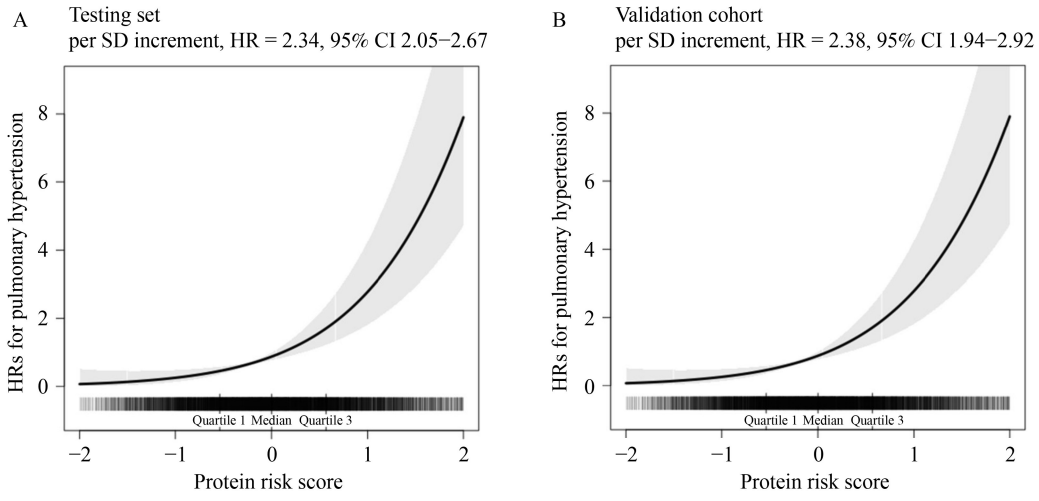

Fig. 1 Association between protein risk score and incident PH risk in (A) the internal testing set and (B) the external validation cohort. Adjusted for age, sex, BMI, smoking status, pulse rate, prevalence of hypertension, diabetes, CVD, CKD, and chronic respiratory disease.

WAP four-disulfide core domain protein 2 (WFDC2), EDN1, lamin-B2 (LMNB2), ADM, and RNA binding protein fox-1 homolog 3 (RBFOX3). These findings were consistently replicated in the external validation cohort (Fig. S3).

Discriminative performance of PH risk prediction models

In the testing set, the PH protein risk score model demonstrated superior discriminative ability for PH risk (C-index = 0.873, 95% CI 0.846–0.900) compared to both the basic demographic model (age and sex; C-index = 0.761, 95% CI 0.726–0.795) and the clinical risk factor model (C-index = 0.843, 95% CI 0.815–0.870) (Table 2). Notably, eight key proteins (ANGPT2, GDF15, NT-proBNP, WFDC2, EDN1, LMNB2, ADM, and RBFOX3) contributed most significantly to this predictive performance, achieving a combined C-index of 0.863 (95% CI 0.835–0.890) (Fig. 3).

Model enhancement analyses revealed that incorporating the protein risk score into the clinical model significantly improved discrimination (C-index increased from 0.843 to 0.881; C-index increase = 0.039, 95% CI 0.001–0.077), while adding clinical factors to the protein model provided minimal improvement (C-index increased from 0.873 to 0.881; C-index increase = 0.008, 95% CI −0.029–0.046) (Table 2). These patterns were consistently replicated in the external validation cohort (Table 2, Fig. 3).

Reclassification performance of PH risk prediction models

The addition of the PH protein risk score to the clinical risk factors model significantly enhanced risk reclassification in the testing set, as evidenced by both continuous NRI improvement (NRI 0.258, 95% CI 0.106–0.336) and IDI improvement (IDI 0.053, 95% CI 0.024–0.089) for 10-year PH risk prediction. These improvements in reclassification performance were consistently observed in the external validation cohort (Table 2).

Sensitivity analysis

Our sensitivity analyses confirmed the robustness of the PH protein risk score across multiple validation approaches. For primary PH prediction, the PH protein risk score demonstrated strong performance in both the testing set (C-index = 0.871, 95% CI 0.831–0.910) and external validation cohort (C-index = 0.875, 95% CI 0.817–0.933), with significant improvement when added to clinical risk factors (Table S8). The PH protein risk score maintained superior discriminative ability after excluding participants with missing protein data (testing set: C-index = 0.897, 95%CI 0.869–0.924; validation cohort: C-index = 0.871, 95% CI 0.815–0.927; Table S9). In addition, 5-fold cross-validation identified 33 proteins for risk score construction, which replicated 73% (22/30) of our primary proteins (Fig. S4). Notably, the 33-protein score showed comparable predictive performance to our primary 30-protein model (development cohort: 0.869 vs. 0.873; validation cohort: 0.881 vs. 0.878; Table S10), demonstrating both the stability of core predictive signatures and the reproducibility of our selection methodology.

Discussion

This large-scale proteomic study identified and validated a 30-protein risk score for incident PH in the general

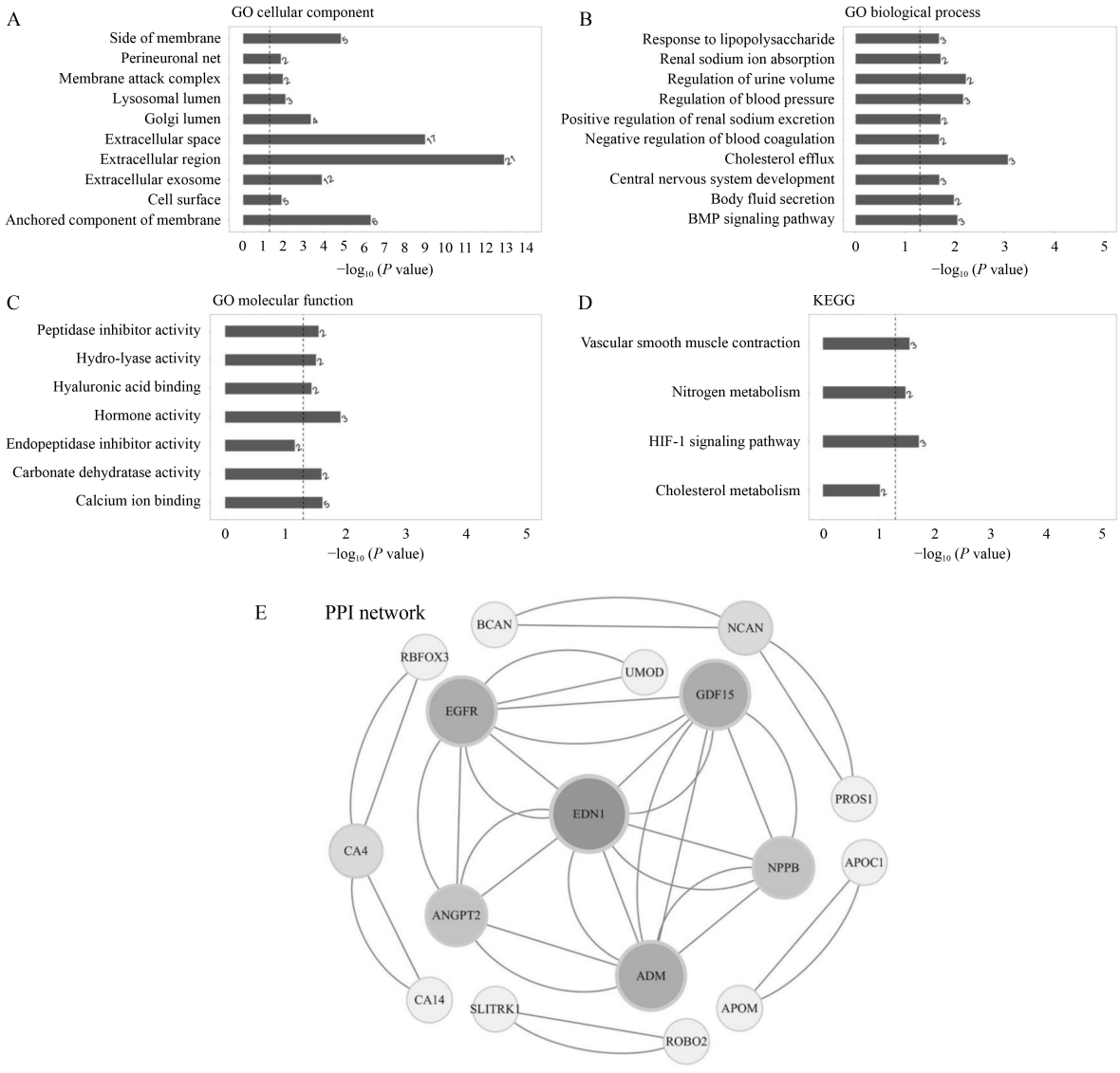

Fig. 2 Functional and pathway enrichment analysis and PPI network of the 30 selected proteins in the PH protein risk score. (A) GO cellular component. (B) GO biological process. (C) GO molecular function. (D) KEGG. (E) PPI network. The number on each bar indicates the count of proteins identified within a specific pathway. In the network visualization, nodes represent proteins, and edges represent protein–protein interactions. Larger nodes denote core proteins within the network.

population, demonstrating superior predictive performance (C-index > 0.87) compared to conventional clinical risk factors. Through comprehensive analyses integrating machine learning, network biology, and MR, we not only developed a robust risk stratification tool but also uncovered novel biological pathways and potential therapeutic targets for PH. Our findings advance the field by addressing critical gaps in current PH prediction models and providing mechanistic insights into disease pathogenesis.

Proteomic profiling outperforms traditional risk assessment

Our study demonstrates that the PH protein risk score exhibits strong discriminative ability (C-index = 0.873) and significantly improves risk reclassification over clinical models (NRI = 0.258). This superior predictive performance (C-index > 0.87) reflects the proteome's unique capacity to integrate inherited risk, environmental exposures, and active pathological processes driving PH development [20,21]. The marginal improvement from adding clinical factors to the PH protein risk score suggests proteomic profiling alone may provide a more efficient risk stratification tool than conventional multi-parameter approaches. Among the eight key predictive proteins, LMNB2 and RBFOX3 emerged as novel PH biomarkers. LMNB2, a nuclear lamina protein that regulates proliferation and DNA methylation, may promote PH through endothelial-to-mesenchymal

Table 2 Performance evaluation of the PH risk prediction model in internal testing and external validation cohorts

Model	C-index	Comparison to reference models		
		C-index increase (95% CI)	10-year risk continuous NRI (95% CI)	10-year risk IDI (95% CI)
Internal testing cohort				
Reference model: basic model (age + sex)	0.761 (0.726, 0.795)			
Clinical risk model[a]	0.843 (0.815, 0.870)	0.082 (0.038, 0.126)	0.331 (0.235, 0.416)	0.014 (0.007, 0.031)
Protein model (age + sex + PH protein risk score)	0.873 (0.846, 0.900)	0.112 (0.068, 0.156)	0.369 (0.279, 0.479)	0.063 (0.032, 0.104)
Reference model: clinical risk model				
Clinical risk model + PH protein risk score	0.881 (0.856, 0.907)	0.039 (0.001, 0.077)	0.258 (0.106, 0.336)	0.053 (0.024, 0.089)
Reference model: PH protein model				
PH protein risk score + clinical risk model	0.881 (0.856, 0.907)	0.008 (−0.029, 0.046)	0.203 (0.075, 0.303)	0.003 (−0.003, 0.024)
External validation cohort				
Reference model: basic model (age + sex)	0.749 (0.693, 0.806)			
Clinical risk model[a]	0.832 (0.784, 0.879)	0.082 (0.009, 0.156)	0.362 (0.182, 0.514)	0.018 (0.008, 0.067)
Protein model (age + sex + PH protein risk score)	0.878 (0.836, 0.919)	0.128 (0.058, 0.198)	0.459 (0.288, 0.604)	0.050 (0.023, 0.107)
Reference model: clinical risk model				
Clinical risk model + PH protein risk score	0.893 (0.857, 0.930)	0.061 (0.002, 0.121)	0.328 (0.097, 0.460)	0.035 (0.005, 0.099)
Reference model: PH protein model				
PH protein risk score + clinical risk model	0.893 (0.857, 0.930)	0.016 (−0.039, 0.071)	0.169 (0.051, 0.423)	0.003 (−0.009, 0.076)

[a]Clinical risk model included age, sex, BMI, smoking status, pulse rate, prevalence of hypertension, diabetes, CVD, CKD, and chronic respiratory disease.
IDI, integrated discrimination improvement; NDI, net reclassification improvement.

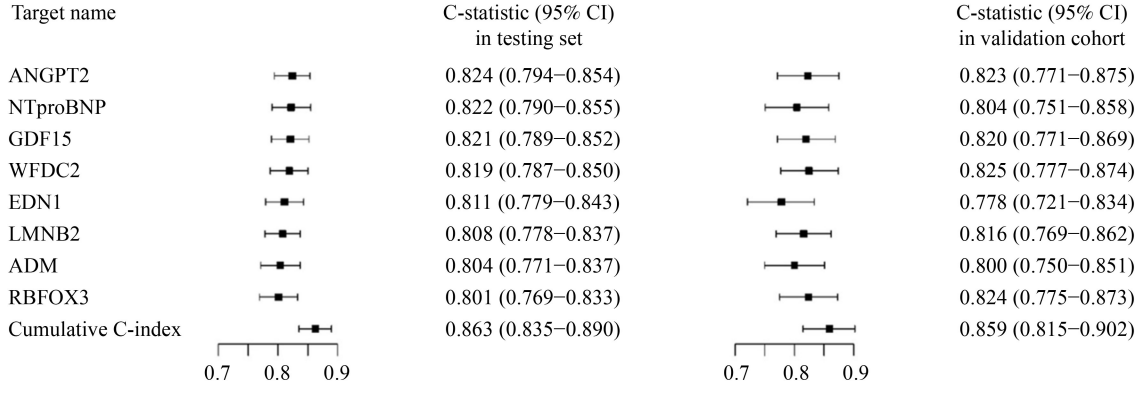

Fig. 3 Individual and cumulative C-index of the eight key proteins for PH risk prediction in (A) the internal testing set and (B) the external validation cohort. The C-index was estimated using a model incorporating age, sex, and each individual protein. The cumulative C-index was derived by calculating the weighted sum of the eight key proteins. ANGPT2, angiopoietin-2; GDF15, growth/differentiation factor 15; NTproBNP, N-terminal prohormone of brain natriuretic peptide; WFDC2, WAP four-disulfide core domain protein 2; EDN1, endothelin-1; LMNB2, lamin-B2; ADM, pro-adrenomedullin; RBFOX3, RNA binding protein fox-1 homolog 3.

transition and vascular remodeling [22,23]. RBFOX3, an RNA splicing regulator, could modulate PH progression via alternative splicing of vascular remodeling genes [24,25]. These findings expand the PH biomarker landscape beyond established candidates (e.g., GDF15 [26], NT-proBNP [27,28], ADM [29,30]), underscoring proteomics' potential for early disease detection.

Causal insights and therapeutic potential

Our MR analysis identified two proteins with causal associations to PH pathogenesis: RGMA (OR = 2.73) and NPC2 (OR = 0.27). RGMA, a vascular patterning molecule currently in clinical trials for multiple sclerosis, promotes vascular smooth muscle cell dedifferentiation and remodeling [31], suggesting its therapeutic potential for PH. Conversely, NPC2's inverse association reveals a previously unrecognized role of cholesterol metabolism in PH, potentially mediated through lipid-driven inflammatory or proliferative pathways [32]. Notably, nine proteins in our risk score represent druggable targets, including both approved therapies (e.g., ANGPT2

inhibitors) and investigational agents, providing immediate opportunities for drug repurposing. While other predictive proteins lacked causal associations, their strong performance as biomarkers warrants further investigation into their roles as disease effectors.

Mechanistic insights and therapeutic implications

Pathway analyses confirmed and extended current understanding of PH pathophysiology, particularly through vascular smooth muscle contraction dysregulation, HIF-1 signaling activation, and calcium homeostasis alterations. The central network position of EDN1 serves dual validation-reinforcing endothelin pathway's known role in vascular remodeling [33] while confirming our proteomic approach's biological relevance.

Advancements and translational implications

Our study significantly advances prior proteomic research in PH by addressing key methodological limitations through establishing generalizability to the general population, achieving superior predictive accuracy via novel protein biomarkers, and providing mechanistic insights for therapeutic target identification. The protein risk score's superior performance over clinical factors highlights the critical need to move beyond conventional risk stratification tools that rely solely on routine clinical variables and lack pathophysiological specificity.

Clinically, this risk score enables transformative opportunities for early intervention by identifying high-risk individuals during the subclinical phase—a crucial advance given the diagnostic delays and poor outcomes characteristic of late-stage PH. Its implementation could guide targeted screening (e.g., echocardiography) and preventive strategies for at-risk subgroups. From a therapeutic perspective, our MR-prioritized targets (RGMA and NPC2) create new research avenues for mechanistic investigation and clinical trials. These collective insights bridge fundamental discovery with clinical application, paving the way for precision medicine approaches in PH management.

Study limitations

Several limitations should be considered when interpreting our findings. First, the exclusive inclusion of White European participants may limit generalizability to other ethnic populations. Second, while the Olink platform provides broad proteomic coverage, it does not encompass all potentially relevant proteins. Third, the statistical robustness of our MR analysis was constrained by the modest number of PH cases in the FinnGen GWAS, necessitating validation in larger datasets. Fourth, the UK Biobank's lack of detailed PH subclassification was partially mitigated by our consistent findings in primary PH sensitivity analyses. Finally, while our integrated clinical and MR analyses identified robust proteomic signatures, experimental validation in preclinical models is needed to confirm their biological roles in PH pathogenesis. Nonetheless, these findings provide clinically meaningful insights into PH pathogenesis and establishe a foundation for future mechanistic and translational research.

In conclusion, this study establishes a protein-based risk score for incident PH that demonstrates superior performance compared to conventional clinical models while providing novel insights into disease mechanisms. By integrating predictive analytics with causal inference and druggability assessments, we present a potential framework for translating proteomic discoveries into clinically useful tools and therapeutic strategies. Future validation studies in prospective cohorts and exploration of early intervention applications will be important next steps. These findings suggest that proteomic approaches may offer promising avenues to improve risk stratification and target discovery in PH, potentially helping to address critical unmet needs in this challenging disease.

Acknowledgements

This research uses data from the UK Biobank. The authors would be willing to thanks the UK Biobank participants. This research was conducted using the UK Biobank Resource under Application Number 73201. This study was supported by the National Key Research and Development Program of China (Nos. 2022YFC2009600, 2022YFC2009605, and 2021YFC2500200); the National Natural Science Foundation of China (Nos. 81973133, 82030022, and 82330020); the President Foundation of Nanfang Hospital, Southern Medical University (No. 2024B029); the Key Technologies R&D Program of Guangdong Province (No. 2023B1111030004); the Guangdong Provincial Clinical Research Center for Kidney Disease (No. 2020B1111170013); and the Program of Introducing Talents of Discipline to Universities, 111 Plan (No. D18005).

Compliance with ethics guidelines

Conflicts of interest Yuanyuan Zhang, Yan Zhang, Sisi Yang, Yu Huang, Yanjun Zhang, Ziliang Ye, Hao Xiang, Xiaoqin Gan, Fan Fan Hou, and Xianhui Qin declared no competing interests for this work.

The study was approved by the North West Multi-Center Research Ethics Committee (11/NW/0382), and the study was performed in accordance with the ethical standards as laid down in the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards. All participants were informed at the initial of the study and signed an informed consent.

Data availability and compliance statement

The authors declare that the acquisition and subsequent use of all data presented in this manuscript fully comply with all relevant local, national, and international laws, regulations, ethical guidelines, and the terms of use associated with the original data sources.

The authors bear full legal responsibility for ensuring the legality of data acquisition and all subsequent uses.

The UK Biobank data are available on application to the UK Biobank, and the analytic methods that support the findings of this study will be available from the corresponding authors on request.

Electronic supplementary material Supplementary material is available in the online version of this article at https://doi.org/10.1007/s11684-025-1183-x and is accessible for authorized users.

References

1. Humbert M, Guignabert C, Bonnet S, Dorfmüller P, Klinger JR, Nicolls MR, Olschewski AJ, Pullamsetti SS, Schermuly RT, Stenmark KR, Rabinovitch M. Pathology and pathobiology of pulmonary hypertension: state of the art and research perspectives. Eur Respir J 2019; 53(1): 1801887
2. Galiè N, Humbert M, Vachiery JL, Gibbs S, Lang I, Torbicki A, Simonneau G, Peacock A, Vonk Noordegraaf A, Beghetti M, Ghofrani A, Gomez Sanchez MA, Hansmann G, Klepetko W, Lancellotti P, Matucci M, McDonagh T, Pierard LA, Trindade PT, Zompatori M, Hoeper M. 2015 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension: The Joint Task Force for the Diagnosis and Treatment of Pulmonary Hypertension of the European Society of Cardiology (ESC) and the European Respiratory Society (ERS): Endorsed by: Association for European Paediatric and Congenital Cardiology (AEPC), International Society for Heart and Lung Transplantation (ISHLT). Eur Heart J 2016; 37(1): 67–119
3. Galiè N, Channick RN, Frantz RP, Grünig E, Jing ZC, Moiseeva O, Preston IR, Pulido T, Safdar Z, Tamura Y, McLaughlin VV. Risk stratification and medical therapy of pulmonary arterial hypertension. Eur Respir J 2019; 53(1): 1801889
4. Bruni C, De Luca G, Lazzaroni MG, Zanatta E, Lepri G, Airò P, Dagna L, Doria A, Matucci-Cerinic M. Screening for pulmonary arterial hypertension in systemic sclerosis: a systematic literature review. Eur J Intern Med 2020; 78: 17–25
5. Qu J, Li M, Wang Y, Duan X, Luo H, Zhao C, Zhan F, Wu Z, Li H, Yang M, Xu J, Wei W, Wu L, Liu Y, You H, Qian J, Yang X, Huang C, Zhao J, Wang Q, Leng X, Tian X, Zhao Y, Zeng X. Predicting the risk of pulmonary arterial hypertension in systemic lupus erythematosus: a Chinese systemic lupus erythematosus treatment and research group cohort study. Arthritis Rheumatol 2021; 73(10): 1847–1855
6. You J, Guo Y, Zhang Y, Kang JJ, Wang LB, Feng JF, Cheng W, Yu JT. Plasma proteomic profiles predict individual future health risk. Nat Commun 2023; 14(1): 7817
7. Williams SA, Kivimaki M, Langenberg C, Hingorani AD, Casas JP, Bouchard C, Jonasson C, Sarzynski MA, Shipley MJ, Alexander L, Ash J, Bauer T, Chadwick J, Datta G, DeLisle RK, Hagar Y, Hinterberg M, Ostroff R, Weiss S, Ganz P, Wareham NJ. Plasma protein patterns as comprehensive indicators of health. Nat Med 2019; 25(12): 1851–1857
8. Ridker PM. Proteomics for the prediction and prevention of atherosclerotic disease. Eur Heart J 2022; 43(16): 1578–1581
9. Carrasco-Zanini J, Pietzner M, Davitte J, Surendran P, Croteau-Chonka DC, Robins C, Torralbo A, Tomlinson C, Grünschläger F, Fitzpatrick N, Ytsma C, Kanno T, Gade S, Freitag D, Ziebell F, Haas S, Denaxas S, Betts JC, Wareham NJ, Hemingway H, Scott RA, Langenberg C. Proteomic signatures improve risk prediction for common and rare diseases. Nat Med 2024; 30(9): 2489–2498
10. Sun BB, Chiou J, Traylor M, Benner C, Hsu YH, Richardson TG, Surendran P, Mahajan A, Robins C, Vasquez-Grinnell SG, Hou L, Kvikstad EM, Burren OS, Davitte J, Ferber KL, Gillies CE, Hedman ÅK, Hu S, Lin T, Mikkilineni R, Pendergrass RK, Pickering C, Prins B, Baird D, Chen CY, Ward LD, Deaton AM, Welsh S, Willis CM, Lehner N, Arnold M, Wörheide MA, Suhre K, Kastenmüller G, Sethi A, Cule M, Raj A; Alnylam Human Genetics; AstraZeneca Genomics Initiative; Biogen Biobank Team; Bristol Myers Squibb; Genentech Human Genetics; GlaxoSmithKline Genomic Sciences; Pfizer Integrative Biology; Population Analytics of Janssen Data Sciences; Regeneron Genetics Center; Burkitt-Gray L, Melamud E, Black MH, Fauman EB, Howson JMM, Kang HM, McCarthy MI, Nioi P, Petrovski S, Scott RA, Smith EN, Szalma S, Waterworth DM, Mitnaul LJ, Szustakowski JD, Gibson BW, Miller MR, Whelan CD. Plasma proteomic associations with genetics and health in the UK Biobank. Nature 2023; 622(7982): 329–338
11. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins R. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 2015; 12(3): e1001779
12. Liu M, Zhang Y, Ye Z, He P, Zhou C, Yang S, Zhang Y, Gan X, Qin X. Enhanced prediction of atrial fibrillation risk using proteomic markers: a comparative analysis with clinical and polygenic risk scores. Heart 2024; 110(21): 1270–1276
13. Zhang Y, Zhang Y, Ye Z, Zhou C, Yang S, Liu M, He P, Gan X, Qin X. Relationship of serum 25-hydroxyvitamin D, obesity with new-onset obstructive sleep apnea. Int J Obes (Lond) 2024; 48(2): 218–223
14. Ye Z, Zhang Y, Zhang Y, Yang S, He P, Liu M, Zhou C, Gan X, Huang Y, Xiang H, Hou FF, Qin X. Large-scale proteomics improve prediction of chronic kidney disease in people with diabetes. Diabetes Care 2024; 47(10): 1757–1763
15. Gan X, Yang S, Zhang Y, Ye Z, Zhang Y, Xiang H, Huang Y, Wu Y, Zhang Y, Qin X. Large-scale plasma proteomics profiles for predicting ischemic stroke risk in the general population. Stroke 2025; 56(2): 456–464
16. Yang S, Ye Z, He P, Zhang Y, Liu M, Zhou C, Zhang Y, Gan X, Huang Y, Xiang H, Qin X. Plasma proteomics for risk prediction of Alzheimer's disease in the general population. Aging Cell 2024; 23(12): e14330
17. Tibshirani R. The lasso method for variable selection in the Cox

model. Stat Med 1997; 16(4): 385–395
18. Eastwood SV, Mathur R, Atkinson M, Brophy S, Sudlow C, Flaig R, de Lusignan S, Allen N, Chaturvedi N. Algorithms for the capture and adjudication of prevalent and incident diabetes in UK Biobank. PLoS One 2016; 11(9): e0162388
19. Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF 3rd, Feldman HI, Kusek JW, Eggers P, Van Lente F, Greene T, Coresh J. CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration). A new equation to estimate glomerular filtration rate. Ann Intern Med 2009; 150(9): 604–612
20. Ferkingstad E, Sulem P, Atlason BA, Sveinbjornsson G, Magnusson MI, Styrmisdottir EL, Gunnarsdottir K, Helgason A, Oddsson A, Halldorsson BV, Jensson BO, Zink F, Halldorsson GH, Masson G, Arnadottir GA, Katrinardottir H, Juliusson K, Magnusson MK, Magnusson OT, Fridriksdottir R, Saevarsdottir S, Gudjonsson SA, Stacey SN, Rognvaldsson S, Eiriksdottir T, Olafsdottir TA, Steinthorsdottir V, Tragante V, Ulfarsson MO, Stefansson H, Jonsdottir I, Holm H, Rafnar T, Melsted P, Saemundsdottir J, Norddahl GL, Lund SH, Gudbjartsson DF, Thorsteinsdottir U, Stefansson K. Large-scale integration of the plasma proteome with genetics and disease. Nat Genet 2021; 53(12): 1712–1721
21. Geyer PE, Kulak NA, Pichler G, Holdt LM, Teupser D, Mann M. Plasma proteome profiling to assess human health and disease. Cell Syst 2016; 2(3): 185–195
22. Li Y, Zhu J, Yu Z, Li H, Jin X. The role of lamin B2 in human diseases. Gene 2023; 870: 147423
23. Kovacic JC, Dimmeler S, Harvey RP, Finkel T, Aikawa E, Krenning G, Baker AH. Endothelial to mesenchymal transition in cardiovascular disease: JACC state-of-the-art review. J Am Coll Cardiol 2019; 73(2): 190–209
24. Duan W, Zhang YP, Hou Z, Huang C, Zhu H, Zhang CQ, Yin Q. Novel insights into NeuN: from neuronal marker to splicing regulator. Mol Neurobiol 2016; 53(3): 1637–1647
25. Liu T, Li W, Lu W, Chen M, Luo M, Zhang C, Li Y, Qin G, Shi D, Xiao B, Qiu H, Yu W, Kang L, Kang T, Huang W, Yu X, Wu X, Deng W. RBFOX3 promotes tumor growth and progression via hTERT signaling and predicts a poor prognosis in hepatocellular carcinoma. Theranostics 2017; 7(12): 3138–3154
26. Nickel N, Kempf T, Tapken H, Tongers J, Laenger F, Lehmann U, Golpon H, Olsson K, Wilkins MR, Gibbs JS, Hoeper MM, Wollert KC. Growth differentiation factor-15 in idiopathic pulmonary arterial hypertension. Am J Respir Crit Care Med 2008; 178(5): 534–541
27. Benza RL, Kanwar MK, Raina A, Scott JV, Zhao CL, Selej M, Elliott CG, Farber HW. Development and validation of an abridged version of the REVEAL 2.0 risk score calculator, REVEAL lite 2, for use in patients with pulmonary arterial Hypertension. Chest 2021; 159(1): 337–346
28. Hoeper MM, Pausch C, Olsson KM, Huscher D, Pittrow D, Grünig E, Staehler G, Vizza CD, Gall H, Distler O, Opitz C, Gibbs JSR, Delcroix M, Ghofrani HA, Park DH, Ewert R, Kaemmerer H, Kabitz HJ, Skowasch D, Behr J, Milger K, Halank M, Wilkens H, Seyfarth HJ, Held M, Dumitrescu D, Tsangaris I, Vonk-Noordegraaf A, Ulrich S, Klose H, Claussen M, Lange TJ, Rosenkranz S. COMPERA 2.0: a refined four-stratum risk assessment model for pulmonary arterial hypertension. Eur Respir J 2022; 60(1): 2102311
29. Kolditz M, Seyfarth HJ, Wilkens H, Ewert R, Bollmann T, Dinter C, Hertel S, Klose H, Opitz C, Grünig E, Höffken G, Halank M. MR-proADM predicts exercise capacity and survival superior to other biomarkers in PH. Lung 2015; 193(6): 901–910
30. Morbach C, Marx A, Kaspar M, Güder G, Brenner S, Feldmann C, Störk S, Vollert JO, Ertl G, Angermann CE. Prognostic potential of midregional pro-adrenomedullin following decompensation for systolic heart failure: comparison with cardiac natriuretic peptides. Eur J Heart Fail 2017; 19(9): 1166–1175
31. Yuan X, Xiao H, Hu Q, Shen G, Qin X. RGMa promotes dedifferentiation of vascular smooth muscle cells into a macrophage-like phenotype *in vivo* and *in vitro*. J Lipid Res 2023; 64(2): 100331
32. Pastore R, Yao L, Hatcher N, Helley M, Brownlees J, Desai R. Deficiency in NPC2 results in disruption of mitochondria-late endosome/lysosomes contact sites and endo-lysosomal lipid dyshomeostasis. Sci Rep 2025; 15(1): 325
33. Shao D, Park JE, Wort SJ. The role of endothelin-1 in the pathogenesis of pulmonary arterial hypertension. Pharmacol Res 2011; 63(6): 504–511