Clinical research protocol

Project name: Multicenter analysis of genomic and metabolic data of neonatal genetic diseases

Leading unit: The Sixth Affiliated Hospital, Sun Yat-sen University

Project leader: Hu Hao

Department: Pediatrics

Research period: September 1, 2022 to December 31, 2025

Project number: EK-YCDX-20220803

NCT No.: none.

Version No.: V1.0

Version Date: August 03, 2022

Catalogue

Abstract

object name	Multicenter analysis of genomic and metabolic data of neonatal genetic diseases
goal of study	(1) Gene sequencing data (138 genes related to 133 common genetic diseases) and tandem mass spectrometry metabolomics data (11 amino acids and 28 acylcarnitines) of about 40,000 newborns from the South China Neonatal Genetic Screening Alliance participating units were collected and collated to complete the database construction of genes and mass spectrometry. (2) A retrospective analysis of the incidence, pathogenic mutations, suspected pathogenic mutations, and the carrying rate of unexplained mutations of 133 common neonatal genetic diseases in 40,000 cases was conducted, and the distribution of high-frequency variation sites in the population was counted, providing a scientific basis for the selection of mutation sites for large-scale genetic screening of neonatal genetic diseases. (3) Protein function prediction and enzyme catalytic activity analysis were performed on high-frequency mutation sites of genetic metabolic diseases (IEM). Characteristic metabolites were analyzed and verified for mutation sites with low enzyme catalytic activity. The feasibility of using protein function artificial intelligence analysis platform and tandem mass spectrometry metabolite data to accurately predict the pathogenicity of genetic metabolic lesions was explored. (4) Explore the use of genome and metabolome big data and machine learning algorithms such as Random forest, Support Vector Machine, Elastic net, Multilayer Perceptron to construct prediction models for common genetic diseases, and strive to achieve accurate diagnosis and prediction of common genetic diseases using simple tandem mass spectrometry metabolome data, and expand the application range of tandem mass spectrometry technology for disease detection.

research design	Cross-Sectional observational study
Research period	September 2022 to December 2025
Leading unit	The Sixth Affiliated Hospital, Sun Yat-sen University
Participating units	South China Neonatal genetic screening Alliance (including cooperation units of 123 hospitals)
Research Consultant	Xin Xiao、Wenhao Zhou
research object	Gene screening data of 40,000 newborns (138 genes related to 133 common genetic diseases) and tandem mass spectrometry data (11 amino acids and 28 acylcarnitines).
Inclusion criteria	(1) Newborns who underwent genetic screening and tandem mass spectrometry at the same time. (2) Age : 0-28 days, gestational age 37-42 weeks.
Excluded criteria	Data that meets any of the following conditions need to be eliminated : (1) Neonatal data with unclear clinical basic information ; (2) Lack of traceability core information data ; (3) The data that the test results cannot be analyzed and interpreted.
sample size	Gene screening and tandem mass spectrometry data of 40,000 newborns.
data collection	(1) Basic information : gender, age, sample type, subject traceability number / ID number, etc. (2) Clinical symptoms, biochemical and imaging data of positive samples. (3) Gene detection results and tandem mass spectrometry results.

	(4) Date of test data, instrument model, reagent type, etc.
Statistical processing	IBM-SPSS 26.0 statistical software was used to analyze the test results. Logistic regression analysis was used to evaluate the correlation between genotype and clinical phenotype. Generalized multifactor dimensionality reduction method was used to analyze gene-gene interaction and genotype-phenotype correlation. Based on tandem mass spectrometry data, machine learning algorithms such as random forest, support vector machine, elastic network, and multi-layer perceptron were used to construct a genotype prediction model. The prediction effect of the model was evaluated by accuracy, balance accuracy, sensitivity, and specificity.
abbreviation	Principal Investigator, PI; tandem mass spectrometry, MS/MS; Inborn errors of metabolism, IEM

1. project basis

1.1 Research background

With the development of China 's social economy, the incidence of infectious diseases and malnutrition diseases has decreased significantly. Birth defects have gradually become a major factor threatening the health of infants and young children, leading to serious disabilities and even death of newborns and infants, such as structural malformations, intellectual development disorders, and language development disorders. It has seriously affected the survival and quality of life of children, brought serious mental and economic burdens to families, and has become a serious public health problem. Preventing and reducing birth defects is one of the important measures to promote the construction of healthy China. The " National Birth Defects Comprehensive Prevention and Control Program " issued by the National Health and Health in 2018 and the " Healthy China Action (2019-2030) " issued by the State Council in 2019 all proposed to comprehensively strengthen the comprehensive prevention and control of birth defects, improve the birth defects prevention and control network, improve the accessibility of comprehensive prevention and control services for birth defects, and carry out in-depth tertiary prevention. Among them, the " Healthy China Initiative (2019-2030) " specifically mentioned the need to pay attention to and strengthen the screening of genetic metabolic diseases and other related birth defects in maternal and child health screening.

Genetic factors are an important cause of birth defects. At present, there are more than 8000 kinds of known single-gene genetic diseases. The overall incidence of genetic diseases is high, the incidence is hidden, and it is easy to cause disability and death. Neonatal disease screening is the last line of life defense for the prevention of birth defects. It is necessary to strengthen the group screening of congenital genetic diseases that seriously endanger the health of newborns. Clinical studies have shown that some birth defects are preventable and treatable, and early diagnosis is the top priority, especially birth defects caused by genetic factors. The clinical phenotypes are

complex and diverse and lack specificity. If they are not diagnosed and treated in time, they often cause disability and death. Through the screening of neonatal genetic diseases, early diagnosis and early intervention of genetic diseases can be achieved, which can avoid or reduce serious complications and nervous system damage. It is a very effective preventive measure to reduce the harm of birth defects.

At present, the commonly used methods of neonatal screening technology include fluorescence immunoassay, enzyme immunoassay and MS / MS technology, but these technologies are all biochemical indicators, and there are some deficiencies : 1 Screening results are affected by gestational age, birth weight, nutrition, medication, environmental differences and other factors, which are prone to false positive results, resulting in maternal and family anxiety ; 2 Screening results are affected by gestational age, birth weight, nutrition, medication, environmental differences and other factors, which are prone to false positive results, resulting in maternal and family anxiety. The false negative rate of some diseases screening was high, such as neonatal intrahepatic cholestasis caused by citrin protein deficiency, polyacyl-CoA dehydrogenase deficiency, ornithine carbamyl transferase deficiency, etc., and their metabolites were usually normal in the neonatal period or when they were not ill, resulting in missed diagnosis. 3 Metabolite detection cannot determine the disease classification of genetic diseases, such as methylmalonic acidemia. There are five common types of methylmalonic acidemia. The severity of different types of symptoms, clinical medication and prognosis are different. (4) Screening diseases can not meet the clinical needs, can not cover some of the higher incidence of metabolic diseases, such as mucopolysaccharidosis, hepatolenticular degeneration and glycogen storage disease. With the improvement of the goal of birth defect prevention in China, neonatal disease screening technology also needs continuous improvement and development to meet the needs of society.

In recent years, gene detection technology has become the gold standard for the diagnosis of most genetic diseases. After more than ten years of iterative updating, the detection accuracy of gene detection technology is getting higher and higher, the detection flux is getting larger and larger, and the detection cost is getting lower and

lower, which makes it possible for gene detection technology to screen newborns. In the past ten years, genetic screening technology has been widely used in many countries such as Europe and the United States, such as preimplantation genetic screening, prenatal screening, neonatal screening, hypertension gene screening, familial hypercholesterolemia gene screening, etc. The results of screening are satisfactory. In 2016, the Chinese Neonatal Genome Project was initiated by the Genetic Counseling Branch of the Chinese Genetic Society and the Children 's Hospital of Fudan University in Shanghai. The program will carry out 100,000 samples of neonatal genetic testing within 5 years, aiming to build a Chinese neonatal genome database, establish neonatal genetic testing standards, promote the industrialization of neonatal genetic testing, and improve the genetic counseling training system. Since the launch of the program, the results have been remarkable. In 2018, Professor Zhou Wenhao 's team of Children 's Hospital Affiliated to Fudan University, head of the Neonatology Group of the Pediatrics Branch of the Chinese Medical Association, took the lead in carrying out a clinical-related whole exome sequencing gene screening research project for newborns. The project has achieved significant results. The data show that genetic testing technology can greatly improve the positive rate of genetic disease screening. In 2020, Professor Zhao Zhengyan, former director of the Pediatrics Branch of the Chinese Medical Association, also conducted neonatal genetic screening research. More than 2500 known pathogenic mutation sites of 135 genes in 75 common genetic diseases were used as detection targets. The data of samples from North China, Northwest, Southwest, Central South and East China showed that the positive detection rate of pathogenic mutations was 2.3 %, and the carrying rate was 26.6 %. Professor Zhao Zhengyan and his team explored a large-scale and large-scale gene screening model based on hot spot mutation detection. In the same year, the ' Beijing neonatal disease screening management method ' clearly included deafness gene screening into the screening content of the neonatal disease screening management method. In early 2021, led by Professor Han Lianshu of Xinhua Hospital Affiliated to Shanghai Jiao Tong University and the head of the new screening center of Shanghai Institute of

Pediatrics, a multi-center neonatal genetic screening research project was launched in China with 8 provincial and municipal tertiary hospitals. 159 genes of 156 common genetic diseases were selected for sequencing screening. At present, nearly 30,000 samples have been analyzed, the positive detection rate is close to 4.0 %, and the carrying rate is 51.98 %. The results show that neonatal genetic screening combined with existing biochemical screening shows obvious advantages, which can detect disease risk earlier, reduce missed diagnosis of conventional new screening, provide detection rate, assist rapid diagnosis of diseases, early intervention, and reduce family and social harm.

In Guangdong Province, our research team took the lead in carrying out a preliminary study of neonatal genetic screening. From June 2019 to April 2020, our team also completed a study of neonatal genetic screening in 1739 cases of Guangzhou Science and Technology Plan Project in Guangdong Province.In this study, high-throughput sequencing technology based on target gene capture was used to detect common genetic disease genes in heel blood samples. The results showed that a total of 158 positive cases were screened out, with a positive rate of 8.81 %. A total of 972 cases carried one or more gene pathogenic variation sites, with a carrying rate of 54.21 %, suggesting that the pathogenic variation of common neonatal genetic diseases has a high incidence and carrying rate in Guangdong, but the study has the problem of insufficient sample size. Compared with the results of traditional neonatal mass spectrometry screening, we found that gene screening has obvious advantages, which effectively reduces the false positive rate and false negative rate of screening, significantly expands the diseases of neonatal genetic disease screening, makes disease diagnosis and typing faster and more accurate, and can specifically design the diseases and genes of gene screening according to regional or ethnic differences.

In recent years, the application of ' artificial intelligence + ' in the medical field has become a hot spot of modern science and technology. The medical industry is highly complex and professional. Artificial intelligence can play a role in many aspects, including medical image recognition, auxiliary diagnosis, drug research and development and other fields. Recently, an artificial intelligence-assisted diagnostic

model was applied to the diagnosis of common pediatric diseases. The results showed that the diagnostic accuracy of 55 common pediatric diseases was close to 90 %. Some diseases such as acute upper respiratory tract infection had a diagnostic accuracy of 95 %, and the average score of the artificial intelligence model was higher than that of the doctor group. Therefore, artificial intelligence can assist early disease diagnosis through a large amount of medical data, improve diagnosis efficiency and reduce misdiagnosis rate, solve the problem of insufficient medical service capacity to a certain extent, and improve the fairness and accessibility of health services. In the field of neonatal disease screening, with the promotion of genetic screening technology and the inclusion of more genetic metabolic diseases in screening, accurate and rapid judgment of multiple disease risks has become an urgent need for clinical front-line. Artificial intelligence has obvious advantages in the interpretation and risk prediction of genetic metabolic disease gene results, which not only significantly reduces the false positive results and saves medical resources ; for areas with weak clinical foundation and urgent need to carry out neonatal genetic metabolic disease gene screening projects, artificial intelligence tools will become a strong support for clinicians, promote the standardization and standardization of genetic metabolic disease diagnosis, and improve the efficiency and accuracy of neonatal genetic metabolic disease gene screening system.

According to the preliminary development of neonatal genetic screening in China, we will retrospectively collect and analyze the genetic screening data and tandem mass spectrometry data of 40,000 newborns in the participating units of the South China Neonatal Genetic Screening Alliance, so as to understand the incidence of neonatal genetic diseases in China and the characteristics of the carrying rate and mutation sites of related gene pathogenic variations, and provide scientific research basis for the feasibility, clinical value and social benefits of large-scale genetic screening. In addition, our previous studies have found that the mutation sites with high carrying rate (more than 1 ‰) in many populations are mutations of unknown clinical significance (VUS). The pathogenic interpretation of such mutations has brought serious confusion to clinicians and brought pending trouble to the subjects.

How to accurately analyze the pathogenicity of a large number of high-frequency VUS mutations has become a key problem to be solved in neonatal genetic screening. Therefore, we intend to use 40,000 cases of gene screening data and tandem mass spectrometry data, and establish a risk assessment model for neonatal genetic metabolic disease gene screening results through protein function artificial intelligence method, in order to solve the pathogenicity problem of high-frequency VUS and provide reliable research basis for clinical practice, so as to quickly and accurately identify high-risk children, reduce the false positive rate of gene screening, and improve the accuracy and system efficiency of neonatal gene screening.

In addition, we will also explore the use of neonatal gene screening data and tandem mass spectrometry data, as well as machine learning algorithms such as random forest, support vector machine, elastic network, and multi-layer perceptron to construct prediction models for common genetic diseases, and strive to achieve accurate diagnosis and prediction of common genetic diseases using simple tandem mass spectrometry metabolomics data, and expand the application range of tandem mass spectrometry technology for disease detection.

1.2 Creativity

This study is a multi-center cooperative study of the South China Neonatal Genetic Screening Alliance. Through the analysis of 40,000 cases of neonatal genetic screening data and MS / MS data, the incidence of common genetic diseases in newborns in China and the carrying rate of related gene pathogenic mutations are deeply understood, and the types of diseases, gene selection, gene screening process, interpretation of gene screening results, follow-up and treatment of screening positive patients and other related contents of neonatal genetic screening are preliminarily clarified, so as to provide scientific research basis for the feasibility, clinical value and social benefits of large-scale genetic screening.

In addition, this study will clarify the correlation between genotypes and metabolic profiles of common genetic diseases in newborns, and explore the feasibility of using protein function artificial intelligence analysis platform and tandem mass spectrometry metabolite data to accurately predict the pathogenicity of

ectopic sites of genetic metabolic lesions through protein function prediction and metabolic level analysis of high-frequency mutation sites of genes, that is, to solve the pathogenicity of mutations with unknown clinical significance. It will also expand the types of disease diagnosis of neonatal MS / MS technology.

This study will also establish a risk assessment model for neonatal genetic metabolic disease gene screening results through artificial intelligence methods. By identifying the pathogenicity of mutation sites, especially those with unclear clinical significance, it can quickly identify high-risk children and reduce false positive rates, thereby improving the accuracy and system efficiency of neonatal genetic screening.

2. Research objective

(1) Gene sequencing data (138 genes related to 133 common genetic diseases) and tandem mass spectrometry metabolomics data (11 amino acids and 28 acylcarnitines) of about 40,000 newborns from the South China Neonatal Genetic Screening Alliance participating units were collected and collated to complete the database construction of genes and mass spectrometry.

(2) A retrospective analysis of the incidence, pathogenic mutations, suspected pathogenic mutations, and the carrying rate of unexplained mutations of 133 common neonatal genetic diseases in 40,000 cases was conducted, and the distribution of high-frequency variation sites in the population was counted, providing a scientific basis for the selection of mutation sites for large-scale genetic screening of neonatal genetic diseases.

(3) Protein function prediction and enzyme catalytic activity analysis were performed on high-frequency mutation sites of genetic metabolic diseases (IEM). Characteristic metabolites were analyzed and verified for mutation sites with low enzyme catalytic activity. The feasibility of using protein function artificial intelligence analysis platform and tandem mass spectrometry metabolite data to accurately predict the pathogenicity of genetic metabolic lesions was explored.

(4) Explore the use of genome and metabolome big data and machine learning algorithms such as Random forest, Support Vector Machine, Elastic net, Multilayer Perceptron to construct prediction models for common genetic diseases, and strive to achieve accurate diagnosis and prediction of common genetic diseases using simple

tandem mass spectrometry metabolome data, and expand the application range of tandem mass spectrometry technology for disease detection.

3. Research scheme and technical route

3.1 Research Design

This study is a multi-center cooperative study of the South China Neonatal Genetic Screening Alliance. The principal investigator (PI) and project leader of this study are Hao Hu, chief physician of pediatrics of the Sixth Affiliated Hospital of Sun Yat-sen University, who plans to include 123 cooperative units of the South China Neonatal Genetic Screening Alliance. In this study, 40,000 neonatal genetic screening data and MS / MS data were retrospectively analyzed through multi-center cooperation. The collection date was from January 2019 to August 2022.

Through the statistical analysis of neonatal genetic screening data (138 genes related to 133 common genetic diseases), the incidence of common genetic diseases in newborns in China, the carrying rate of pathogenic variation and the high-frequency variation sites of the population were clarified, and the epidemiological characteristics of newborns in China were studied.

Through the statistical analysis of neonatal genetic screening data and MS / MS metabolomics data (11 amino acids and 28 acylcarnitines), the correlation between gene and metabolism will be explored, and the pathogenicity of high-frequency VUS mutation sites will be identified by using protein function artificial intelligence analysis platform and tandem mass spectrometry metabolite data.

The prediction model of common genetic diseases is constructed by using machine learning algorithms such as random forest, support vector machine, elastic network and multi-layer perceptron, so as to realize the accurate diagnosis of common genetic diseases through tandem mass spectrometry metabolomics data, and expand 2-3 kinds of diseases that can be detected by MS / MS technology.

3.1 Sample size

This study plans to collect genetic screening data (138 genes related to 133 common genetic diseases) and tandem mass spectrometry metabolomics data (11

amino acids and 28 acylcarnitines) of about 40,000 newborns from January 2019 to August 2022 in 123 cooperative units of the South China Neonatal Genetic Screening Alliance.

3.3 data collection

In the course of the study, no random or any treatment driven by the research program will be performed or provided to the subjects. The researchers will view the patient 's medical history and laboratory reports, and determine the patient 's eligibility according to the inclusion and exclusion criteria.

(1) The types of diseases detected by genetic screening : including 133 kinds of hereditary diseases with high incidence in neonates. The types of diseases include genetic metabolic diseases such as amino acid metabolic diseases, organic acid metabolic diseases and fatty acid metabolic diseases within the scope of all neonatal tandem mass spectrometry screening, as well as other genetic diseases that cannot be screened by tandem mass spectrometry but have high incidence and early detection of great clinical significance and social benefits for children and families, such as lysosomal diseases, glycogen accumulation diseases, hepatolenticular degeneration, hereditary deafness, β-thalassemia, etc.

(2) In this study, the data analysis of genes and the interpretation of test reports are based on the implementation of industry consensus or guidelines for genetic testing of genetic diseases such as ' Discussion on the standardized consensus of the whole process of clinical detection of genetic diseases second-generation sequencing '. The analysis and interpretation of test data refer to the ' ACMG genetic variation classification criteria and guidelines ' issued by the American Society of Medical Genetics and Genomics (ACMG) in conjunction with the Molecular Pathology Society (AMP) and the domestic ' genetic variation classification criteria and guidelines (2017) '. The variation was divided into five grades : pathogenic (P, pathogenic possibility > 0.99), suspected pathogenic (LP, pathogenic possibility 0.95-0.99), unknown clinical significance (VUS, pathogenic possibility 0.01-0.90), suspected benign (LB, pathogenic possibility

0.001-0.01), benign (B, pathogenic possibility < 0.001). The results of gene detection in this study only classified pathogenic and possible pathogenic mutations as positive mutations, and homozygotes and compound heterozygotes of autosomal recessive inheritance, single heterozygotes and homozygotes of autosomal dominant inheritance, and hemizygotes of X-linked inheritance were classified as disease positive. The unclear clinical significance is not directly used as the basis for clinical diagnosis. It is necessary to combine the clinical, if necessary, to carry out the corresponding family member verification, auxiliary examination, clinical follow-up.

(3) MS / MS data collection and standardization : The collected data include (1) laboratory background information including instrument type, reagent type, quality control material type, laboratory quality control rules, filter paper type, normal population range ; (2) Quality control data included quality control number, quality control type, quality control batch number, amino acid internal standard batch number, acylcarnitine internal standard batch number, factory time, experimental time, and test value of each quality control analyte ; (3) Screening test data include screening number, gestational age, breastfeeding method, native place (province, city), newborn gender, birth date, birth weight, sample number, blood collection time, etc. ; (4) The data of confirmed cases included screening number, confirmed disease, blood ammonia test, blood gas analysis, blood routine, liver function test, urine GC-MS test, imaging examination, gene detection information and other tests. The fluctuation range of the detection concentration of the corresponding batch of quality control samples was within the target value ±3 s. Finally, the data were collected and sorted into the negative database and the positive database. In order to eliminate the influence of regional differences, the series detection index data were standardized.

(4) By analyzing the correlation between large sample gene data and tandem mass spectrometry data and the correction of clinical prognosis, a neonatal genetic disease gene screening database and an artificial intelligence risk assessment

model were established. Construction and selection of modeling indicators, select core indicators through information gain and correlation coefficient, and construct combined features ; in the model training stage, the gene screening data is divided into training set and test set at a ratio of 8 ： 2, and the ensemble learning models such as random forest, gradient boosting tree algorithm and artificial neural network are trained to select the best model. Model evaluation, the model selection process first meets the detection rate of 1, and then compares the false positive rate to select the optimal training model. Model verification, using a new batch of big data samples, through model calculation, output high-risk and low-risk sample lists, and compare and analyze with the original screening results.

3.4 Inclusion criteria

(1) Newborns who underwent genetic screening and tandem mass spectrometry at the same time.

(2) Age : 0-28 days, gestational age 37-42 weeks.

3.5 Excluded criteria

Data that meets any of the following conditions need to be eliminated :

(1) Neonatal data with unclear clinical basic information ;

(2) Lack of traceability core information data ;

(3) The data that the test results cannot be analyzed and interpreted.

3.6 Observation items

(1) Positive detection rate of common genetic diseases in neonates : the results of genetic testing of enrolled neonates showed that the proportion of homozygous and compound heterozygotes of autosomal recessive inheritance, single heterozygote and homozygous of autosomal dominant inheritance, and hemizygote of X-linked inheritance accounted for the total number of people tested.

(2) The corresponding pathogenic genes and pathogenic variation carrying rate of common neonatal genetic diseases : that is, the proportion of the total number of occurrences of pathogenic variation, suspected pathogenic variation and

unknown clinical significance variation in the genetic test results of the enrolled neonates according to the ' ACMG genetic variation classification criteria and guidelines ' and ' genetic variation classification criteria and guidelines (2017) '.

(3) The carrying rate of high-frequency mutation sites in the population : that is, the mutation sites with a population carrying rate greater than 1 ‰.

(4) Establish a complete protein structure and catalytic activity analysis process, and develop artificial intelligence protein analysis tools for pathogenicity prediction of mutation sites.

(5) Preliminary identification of the pathogenicity of high-frequency VUS mutation sites : The protein structure and enzyme catalytic activity characteristics of high-frequency VUS mutation sites were predicted by artificial intelligence tools, and the metabolic spectrum differences between positive, carrier and negative samples of high-frequency VUS mutation sites were clarified by combining with tandem mass spectrometry metabolite data, which provided a basis for the pathogenicity judgment of VUS mutation.

(6) Gene-metabolism correlation : machine learning algorithms such as random forest, support vector machine, elastic network and multi-layer perceptron are used to construct prediction models of common genetic diseases, so as to realize accurate diagnosis of common genetic diseases through tandem mass spectrometry metabolomics data, and expand 2-3 diseases that can be diagnosed by tandem mass spectrometry.

3.7 Statistical processing

IBM-SPSS 26.0 statistical software and R language (Version 4.1.0) were used for statistical analysis of the test results, and the positive rate of neonatal genetic disease screening (common positive disease cases, common positive sites) was counted. The carrying rate of common mutation sites in newborns was counted. The measurement data were expressed as mean ±standard deviation ($\bar{x}$ ±s), and t test was used for comparison between groups. The count variables were expressed as values or rates (%), and the comparison between groups was tested. Unconditional logistic regression analysis was used to evaluate the correlation between genotype and

clinical phenotype. Generalized multifactor dimensionality reduction (GMDR) was used to analyze gene-gene interaction and genotype-phenotype correlation. Based on tandem mass spectrometry data, machine learning algorithms such as random forest, support vector machine, elastic network, and multi-layer perceptron were used to construct genotype prediction models. Accuracy, Balanced accuracy, sensitivity, and specificity were used to evaluate the prediction effect of the model.

4. Ethics in clinical research

4.1 Data source

The gene sequencing data and MS / MS metabolic data of 40,000 newborns were from 123 cooperative units of the South China Neonatal Gene Screening Alliance.In this study, the data table established by Microsoft Excel was used. The neonatal gene data and tandem mass spectrometry of the multi-center cooperative units were transmitted through the Excel data table. Effective measures will be taken to strictly record, clean and check the data. Multi-centers ensure the authenticity, accuracy and completeness of the neonatal gene sequencing data and MS / MS metabolic data provided, and all data and test reports can be traced. In addition, the data management, pay attention to the confidentiality of the data, to ensure the privacy of patients and their families.

4.2 Informed consent

This study is a retrospective study. Subjects have signed informed consent from parents or guardians when doing neonatal MS / MS metabolic disease screening or genetic screening. The informed consent form clearly states that the test data can be used for scientific research after removing personal privacy information. Therefore, the application for exemption from informed consent.

4.3 Benefits of participating in research

There is no direct economic benefit for all the test subjects included in this study, but for the positive children included in this study, free first-generation sequencing verification is provided, and professional genetic counseling and clinical treatment advice are provided to parents by pediatric clinicians with the consent of the parents of the children.

4.4 Privacy protection measures

All the data of the subjects during the study period will be entered into the computer for confidential storage and analysis. If necessary, the relevant institutions may review the records to confirm the authenticity, accuracy and integrity of the data. The data obtained from the study may also be published in academic journals, but the names of the subjects will not be published, and the privacy of the subjects will be kept confidential.

All selected populations do not involve special populations, and patient privacy information is strictly protected.

4.5 Data Recording

The results of this study will be recorded in human genetic resources in China.

5. Research progress

September 01, 2022 - December 31, 2022: Prepare programs and projects, sign cooperation agreements with partners.

January 01,2023 to December 31,2024: (1) Data processing: Collect gene sequencing data and MS / MS data, and organize the data to establish a database of genes and mass spectrometry. (2) Data analysis : Gene data and mass spectrometry were analyzed by statistical software. Data, get the target statistical results. (3) Establishment of protein structure and catalytic activity analysis process, development of artificial Intelligent protein function analysis tool. (4) Completed the preliminary identification of the pathogenicity of high-frequency VUS mutation sites.

January 01, 2025 - December 31, 2025: (1) To study the correlation between gene and metabolism, and to construct a string of common genetic diseases. Mass spectrometry prediction model, and increase the use of MS / MS technology for diagnosis The disease, and to verify. (2) Write the conclusion report, complete the paper, submission and publication work.

6. Reference

[1] Report on prevention and treatment of birth defects in China (2012) [EB/OL].(2012-9-12)[2018-12-12]. http://www.moh.gov.cn/wsb/pxwfb/201209/55840.

[2] Wang Q, Xiang J, Sun J, et al. Nationwide population genetic screening improves outcomes of newborn screening for hearing loss in China[J]. Genet Med, 2019;21(10):2231-2238. doi:10.1038/s41436-019-0481-6.

[3] Zhong K, Wang W, He F, Wang Z. The status of neonatal screening in China,2013[J]. J Med Screen, 2016, 23(2):59-61. doi:10.1177/0969141315597715.

[4] Deng K, He C, Zhu J, et al. Incidence of congenital hypothyroidism in China: data from the national newborn screening program,2013-2015[J]. J Pediatr Endocrinol Metab, 2018,31(6):601-608. doi:10.1515/jpem-2017-0361.

[5] Li LH, Wu WC, Li N, et al. Full-Term Neonatal Ophthalmic Screening in China: A Review of 4-Year Outcomes[J]. Ophthalmic Surg Lasers Imaging Retina, 2017;48(12):983-992. doi:10.3928/23258160-20171130-05.

[6] Dai P, Huang LH, Wang GJ, et al. Concurrent Hearing and Genetic Screening of 180,469 Neonates with Follow-up in Beijing, China[J]. Am J Hum Genet, 2019;105(4):803-812. doi:10.1016/j.ajhg.2019.09.003.

[7] Lyu K, Xiong Y, Yu H, et al. Screening of common deafness gene mutations in 17 000 Chinese newborns from Chengdu based on microarray analysis[J]. Zhonghua Yi Xue Yi Chuan Xue Za Zhi, 2014;31(5):547-552. doi:10.3760/cma.j.issn.1003-9406.2014.05.001.

[8] Hao Z, Fu D, Ming Y, et al. Large scale newborn deafness genetic screening of 142,417 neonates in Wuhan, China[J]. PLoS One, 2018, 13(4):e0195740. doi:10.1371/journal.pone.0195740.

[9] Wang Q J, Shen Y P, Wu L Q, et al. Standards and guidelines for the interpretation of sequence variants[J]. Sci Sin Vitae, 2017, 47:1-21. doi:10.1360/N052017-00099.

[10] Currier RJ. Single-Gene Sequencing in Newborn Screening: Success, Challenge, Hope[J]. Hastings Cent Rep, 2018, 48 (2):S37-S38. doi:10.1002/hast.883.

[11] Ceyhan-Birsoy O, Murry JB, Machini K, et al. Interpretation of Genomic Sequencing Results in Healthy and Ill Newborns: Results from the BabySeq Project[J]. Am J Hum Genet, 2019, 104(1):76-93. doi:10.1016/j.ajhg.2018.11.016.

[12] Momosaki K, Kido J, Yoshida S, et al. Newborn screening for Pompe disease in Japan: report and literature review of mutations in the GAA gene in Japanese and Asian patients[J]. J Hum Genet, 2019, 64(8):741-755. doi:10.1038/s10038-019-0603-7.

[13] Verma IC, Puri RD. Global burden of genetic disease and the role of genetic screening[J]. Semin Fetal Neonatal Med, 2015, 20(5):354-363. doi:10.1016/j.siny.2015.07.002.

[14] Boardman FK, Sadler C, Young PJ. Newborn genetic screening for spinal muscular atrophy in the UK: The views of the general population[J]. Mol Genet Genomic Med, 2018 , 6(1):99-108. doi: 10.1002/mgg3.353. Epub 2017 Nov 23.

[15] Zou Y, Dai QQ, Tao WJ, et al. Suspension array-based deafness genetic screening in 53,033 Chinese newborns identifies high prevalence of 109 G>A in GJB2[J]. Int J Pediatr Otorhinolaryngol, 2019, 126:109630. doi:10.1016/j.ijporl.2019.109630.