

RESEARCH



# Exploring the application of machine learning to identify the correlations between phthalate esters and disease: enhancing nursing assessments

Hao-Ting Wu<sup>1</sup>, Chien-Chang Liao<sup>2</sup>, Chiung-Fang Peng<sup>3</sup>, Tso-Ying Lee<sup>4\*</sup> and Pei-Hung Liao<sup>5\*</sup> 

## Abstract

**Background:** Health risks associated with phthalate esters depend on exposure level, individual sensitivities, and other contributing factors.

**Purpose:** This study employed artificial intelligence algorithms while applying data mining techniques to identify correlations between phthalate esters [di(2-ethylhexyl) phthalate, DEHP], lifestyle factors, and disease outcomes.

**Methods:** We conducted exploratory analysis using demographic and laboratory data collected from the Taiwan Biobank. The study developed a prediction model to examine the relationship between phthalate esters and the risk of developing certain diseases based on various artificial intelligence algorithms, including logistic regression, artificial neural networks, and Bayesian networks.

**Results:** The results indicate that phthalate esters exhibited a greater impact on bone and joint issues than heart problems. We observed that DEHP metabolites, such as mono(2-carboxymethylhexyl) phthalate, mono-n-butyl phthalate, and monoethylphthalate, leave higher residue in females than in males, with statistically significant differences. Monoethylphthalate levels were lower in individuals who exercised regularly than those who did not, indicating statistically significant differences.

**Conclusions:** This study's findings can serve as a valuable reference for clinical nursing assessments regarding diseases related to osteoporosis, arthritis, and musculoskeletal pain. Medical professionals can enhance care quality by considering factors beyond patients' essential physical assessment items.

Trial Registration: This study was registered under NCT05892029 on May 5, 2023, retrospectively.

**Keywords:** Environmental toxic substances, Phthalate esters, Machine learning, Nursing assessments, Prediction model

## Introduction

Primarily composed of phthalates [di(2-ethylhexyl) phthalate, DEHP], phthalate esters (PAEs) are commonly used in polyvinyl chloride products, representing the

most prevalent environmental hormones in daily life [1]. These compounds can enter the human body through inhalation, ingestion, drinking, or skin contact. Their effects are impacted by dosage, exposure time, route of entry, and processing methods [2, 3]. Environmental factors, genetic factors, disease vectors, and lifestyle choices affect individual health, requiring comprehensive observations, measurements, and tracking [3]. While PAE test data are typically stored in hospital databases, a lack of standardization across hospitals results in test

\*Correspondence: tsoyinglee@gmail.com; peihung@ntunhs.edu.tw

<sup>4</sup> Nursing Reserach center & School of Nursing, Taipei Medical University Hospital & Taipei Medical University, Taipei, Taiwan

<sup>5</sup> School of Nursing, National Taipei University of Nursing and Health Sciences, No. 365, Ming-Te Road, Peitou District, Taipei 112, Taiwan  
Full list of author information is available at the end of the article

item variation; there is often no integration between PAE test results and lifestyle questionnaires [3, 4]. Thus, a vast amount of relevant data remains unexplored. However, recent studies use data mining techniques to analyze National Health Insurance data while developing disease risk assessment models to prevent disease occurrence [4].

Environmental hormones are present in various materials and chemical substances to which people are commonly exposed (e.g., food, clothing, housing materials, and transportation), while plastic products are particularly affected [5]. Health risks associated with PAEs are caused by various factors, including exposure levels, individual sensitivities, and other promoting factors [4]. DEHP is a chemical compound that enhances plasticity and ductility within plastic products. It is toxic to animals, taking one to two days for the human body to digest. Long-term exposure to high amounts of DEHP can lead to environmental hormone changes, disrupting the endocrine system [5, 6]. Researchers have recently employed risk prediction models to investigate this issue [6].

Data mining and deep learning techniques are valuable tools for examining disease risk factors within the vast repositories of electronic medical records. These tools can foster early disease risk assessment methods. Researchers often use algorithms to analyze large datasets to predict disease risk [7, 8]. For instance, Li et al. constructed predictors for the negative effects of environmental hormones on various systems while identifying the mechanisms by which environmental particulate matter can trigger toxic effects [9]. Liu et al. developed a prediction model for human bladder cancer that is induced by aromatic amines. Their study identified carcinogenesis mechanisms while screening alternative aromatic amine antioxidant molecules with lower carcinogenicity for bladder cancer [10, 11]. Most prior studies have adopted adverse outcome pathways (AOPs) to identify the effects of toxic substances, while machine learning models have seldom been employed [9].

Phthalate esters are found in several daily substances including pesticides, plasticizers, pharmaceuticals, personal care products, food, and food packaging. Bis(2-ethylhexyl) phthalate (DEHP) is often added to PVC products in industrial applications to increase plasticity and ductility [11]. However, it is toxic to animals. Although the body can metabolize it over approximately one or two days, it can have environmental and hormonal effects that interfere with the endocrine system under long-term exposure to large amounts [12]. However, the public's understanding of plasticizers is generally insufficient; the ability of healthcare personnel to conduct further evaluation and screening is relatively weak. This study investigated the correlation between physiological

test data, environmental hormone plasticizer content, lifestyle habits, and disease using artificial intelligence algorithms. The objective was to establish a prediction model for high-risk diseases associated with environmental hormone plasticizers. The findings can enhance clinically relevant nursing assessments and the early detection of the impact of individual lifestyle on disease and early clinical symptoms.

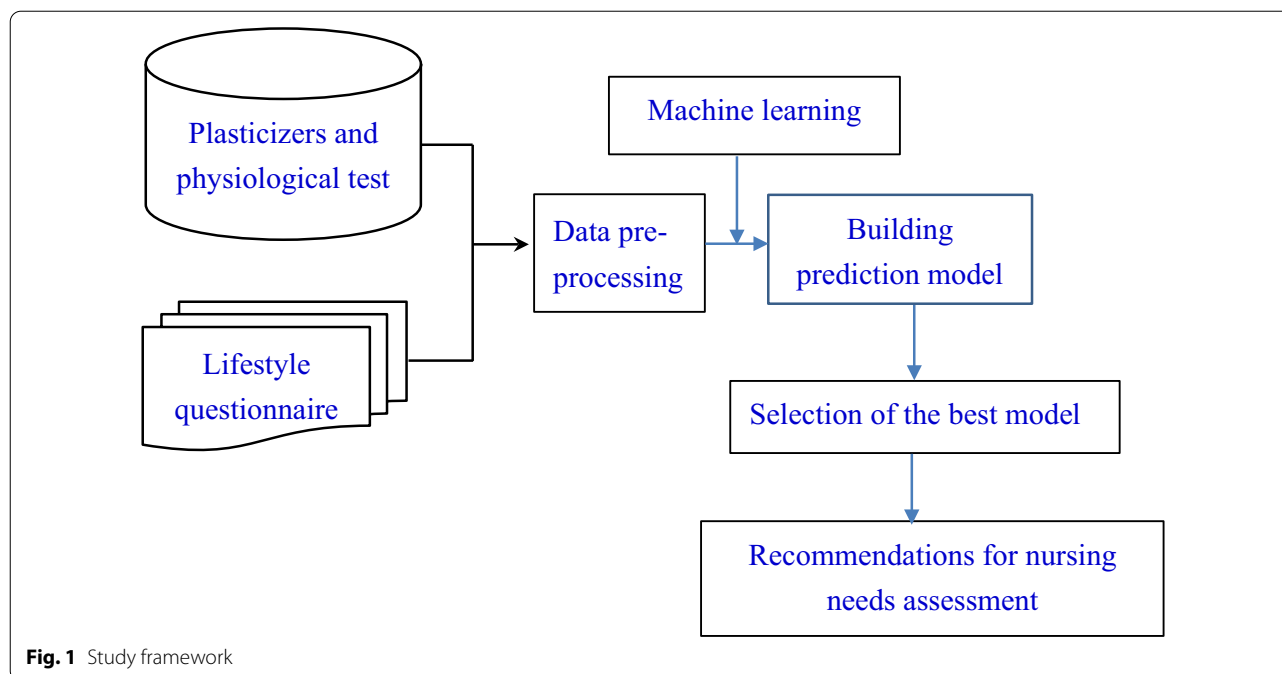
Current research indicates that environmental hormones can negatively affect the body. This study analyzed case tracking data obtained from the Taiwan Biobank based on the following components. (1) General participant questionnaire content: (a) basic personal information, (b) individual health behavior, (f) female health issues, comprising three units. (2) General participant physical examination contents: body mass index (BMI), body fat percentage, waist and hip circumference, waist-to-hip circumference ratio, blood pressure, heart rate, lung function, and bone mineral density. (3) General participant blood and urine analysis items: hematology tests, serology tests, hepatobiliary function tests, renal function tests, and urine tests. (4) Plasticizer content in urine (MEHP, MEOHP, MEHHP, MECPP, MCMHP, MBzP, MnBP, MiBP, MEP, MMP, and MiNP: 11 plasticizer metabolites). The analysis sought to determine the correlation between plasticizer levels in the body and lifestyle factors, genetic predisposition, and disease outcomes. This study established a prediction model for the risk of plasticizers contributing to disease development, facilitating early detection of lifestyle-related disease impacts and early clinical symptoms. The objective is to enable early prevention of related diseases while ensuring the safety of the living environment (Fig. 1).

## Methods

### Study design

This study conducted exploratory research and received approval from the Institutional Review Board (IRB) and the Taiwan Biobank prior to data collection (approval number: TWBR11007-06). Taiwan Biobank data were voluntarily provided by subjects participating in examinations across Taiwan. A researcher cannot file an application to the Taiwan Biobank until a study is reviewed and approved by the research institution's IRB. Delinked data where personal information has been removed can be obtained after a study is reviewed and approved and payment is made. All data will be destroyed within two years after a study is published.

The dataset comprised information collected from participants between 30 and 70 years old who underwent testing for PAEs between 2016 and 2022, totaling 1,337 observations. This data comprised the following components. (1) Questionnaire responses, including basic



personal information, individual health behaviors, and female health issues. (2) Physical examination results, including body mass index (BMI), body fat percentage, waist and hip circumference, waist-to-hip ratio, blood pressure, heart rate, pulmonary function, and bone mineral density. (3) Blood and urine analysis results from blood tests, serology tests, hepatobiliary function tests, renal function tests, and urinalysis. (4) Data regarding PAE content in urine.

Using statistical analysis, the events per variable (EPV) method is often employed to determine sample sizes for new model development. EPV requires at least ten predictive variables for each event per variable. Thus, the model receives at least 10 instances of training for each variable [13]. Osteoarthritis-related variables were determined to be the most sufficient within the Biobank dataset. Variables regarding other diseases were reinforced with mean values.

#### Patient and public involvement

Since this study primarily involved database analysis, patients and the public were not involved in the study design, implementation, reporting, or dissemination plan. Several contributing studies incorporated patient and community stakeholder involvement in the design and dissemination of findings.

#### Analysis tools

This study used IBM SPSS 25.0 for Windows and IBM SPSS (English version) Modeler 18.2 for data processing

and analysis. IBB SPSS 25.0 for Windows is a statistical software program that analyzes the correlation and risk levels among critical factors and related diseases based on Chi-square tests, t-tests, and odds ratio (OR). IBM SPSS Modeler 18.3 comprises data mining and prediction analysis software. This study employed data mining algorithms such as Bayesian and neural networks to develop a prediction model while visually illustrating model accuracy [8].

The CRISP-DM procedural model organizes the life cycle of data mining into six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Since most data from the Taiwan Biobank originate from physical examinations from initial subjects, not all research examines the same items; some of these data are missing at random (MAR). The occurrence of missing data is associated with observed data, independent of unobserved data. Therefore, this study employed the hot deck imputation method to account for missing data. This method assumes that while the entire dataset of a respondent's basic information, only Respondent L (for example) had omitted their weight. It is possible to refer to other characteristics Respondent L provided, e.g., height, age, and gender, and determine the respondent most similar, filling in such person's data. The phases are detailed below [14, 15]:

### **Unconnected data**

Taiwan Biobank data was the primary data source.

### **Data pre-processing**

Data were screened before analysis. Incomplete entries were excluded. The remaining data were standardized to minimize noise and ensure prediction model accuracy.

### **Screening research variables**

High-risk factors were screened based on the relevant literature. Key discriminant factors were selected from individual datasets from the Taiwan Biobank and integrated as input variables in this study.

### **Establishing a classification model**

We established a prediction model for disease risk as follows. First, the IBM SPSS Modeler 18.2 machine learning algorithm was used to examine the impact of individual lifestyle factors on the concentration of environmental hormones in human bodies. We employed binary logistic regression and OR to predict the risk factors of PAE-induced disease. We employed a Chi-square test to analyze the correlation between PAEs and related diseases. Finally, we used IBM SPSS Modeler 18.0 artificial neural and Bayesian networks to develop the prediction model and weigh the predictors.

### **Model selection**

The model with the highest accuracy rate was selected among several models tested.

### **Discussing the results and suggestions for improvement**

The analysis results were synthesized; recommendations for enhancing the nursing assessment system were proposed.

Machine learning can be divided into two main categories. (1) Supervised learning: meaning there is a target variable. Individuals must explore the relationship between feature variables and the target variable to learn and optimize the algorithm under target variable supervision. (2) Unsupervised learning: involving the absence of a target variable. This prompts individuals to recognize the inherent patterns and features between variables based on the data itself [16, 17]. Machine learning is an artificial intelligence approach that can be used to address problems encountered based on artificial intelligence processes. Several open-source machine and deep learning tools use modern artificial intelligence technology (e.g., Theano, TensorFlow, MXNet, CNTKt, Keras). With a vast amount of diverse clinical data related to human disease, these expert

systems adapt the logic and rules of medical experts to diagnose disease. These systems can compare patients' problems with regulations and facts obtained through machine learning, determining the most suitable diagnosis. Based on relevant literature, we obtained information on the following: risk factors highly correlated with plasticizers, physical assessment questionnaire content for individual cases, physiological test data, and plasticizer metabolite content in urine. Since Chinese people do not know much about environmental hormone detection, the resources available for this study are limited. Thus, basic machine learning algorithms were used. Based on variable attributes, this study used logistic regression, artificial neural networks, Bayesian network algorithms, and XG-boosts algorithms for prediction and data analysis. Definitions of the four machine learning algorithms are presented in Fig. 2 [9, 11, 18].

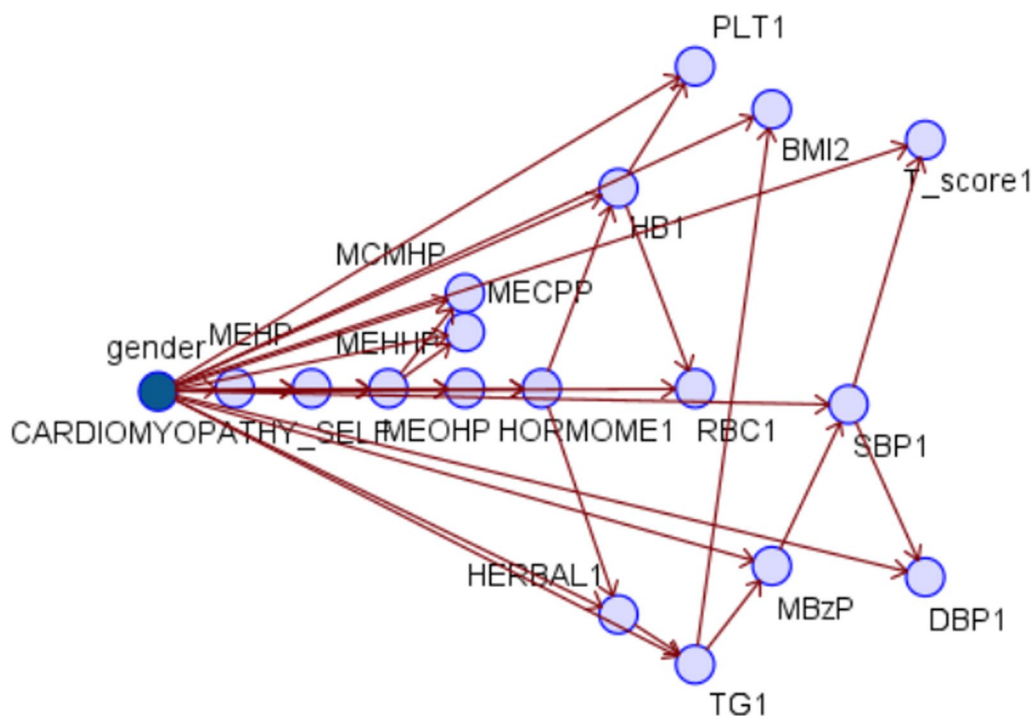
### **Bayesian networks**

The Bayesian network algorithm model is rooted in classical mathematical theory. It provides a solid mathematical foundation while exhibiting stable classification efficiency. It performs well using small-scale datasets and can handle multi-classification tasks individually. This model is suitable for incremental training (i.e., it can train on new samples in real-time) and is not sensitive to missing data. Its algorithm is relatively simple and facilitates the straightforward interpretation of results [16, 19].

In a Bayesian network, the nodes in the directed acyclic graph indicate random variables. These are classified as observable, hidden, and unknown parameters, among others. Arrows connect the variables or propositions presumed to have causal relationships or lack conditional independence. A single arrow connecting two nodes defines one as the parent and the other as the child, each with a conditional probability value. Bayesian networks are generated by drawing random variables in a directed graph, depending on conditional independent relationships [9].

### **Artificial neural networks**

An artificial neural network is a software and hardware-based computing system that uses several connected artificial neurons to imitate a biological neural network. Through continual training and learning processes, it accumulates experiences to obtain optimal results. Its software is trained to provide the most appropriate analysis, classification, and prediction by adjusting variable settings and weight values. Through training with sample data, ANNs can develop a system model that depicts the underlying patterns and relationships within data. Once



**Fig. 2** Bayesian network: cardiomyopathies as an example

trained, this model can estimate, predict, diagnose, and select new or unseen data. Comprising multiple artificial cells known as neurons, artificial neurons, or processing units, each processing unit's output becomes the input for other units [9, 10].

### XGBoost (eXtreme gradient boosting)

XGBoost (eXtreme Gradient Boosting) is an advanced gradient-boosting algorithm. It contains the keyword 'Boosting' to indicate that it is a boosting integration algorithm. Its premise is to combine hundreds or thousands of tree models with high accuracy, generating new models through a continuous iteration tree [9–11, 18].

## Results

### Basic demographics and relevant variable analysis

This study collected data regarding individuals aged 30 to 70 who underwent testing for PAEs from January 1, 2016, to December 31, 2020. Combing the Taiwan Biobank, we identified a total of 1337 cases. The data exported from the biobank included: (1) basic personal information, (2) individual health behavior information, (3) information on female health issues, (4) physical examination results, (5) blood and urine analysis results, and (6) data regarding PAE urine content. We conducted a descriptive analysis of various demographic and health-related factors, including age, gender, marital status, education

level, drinking habits, exercise habits, living area, job classification, body mass index (BMI), body fat percentage, and bone density of the study subjects. Among the 1337 respondents, the average age was 48.98 years. Male respondents (51.8%) outnumbered female respondents. Furthermore, 75.1% of respondents were married, 9.4% reported drinking habits, 57.4% did not exercise regularly, 15.3% worked in the manufacturing industry, 47.4% were classified as overweight based on their BMI, 45.2% had a body fat percentage higher than normal, and 16.8% exhibited bone density below 2.5. Table 1 presents the analysis results.

The basic information indicates that gender, educational level, exercise habits, BMI, and body fat percentage were relatively evenly distributed. We analyzed these variables and found a significant difference between gender and experience habits. Table 2 presents the gender differences in PAE metabolites in urine within the 1,337 analyzed samples (693 males and 644 females). Results from the independent t-test conducted on the female urine samples indicated higher levels of the metabolites mono (2-carboxymethylhexyl) phthalate (MCMHP) ( $t = -2.35$ ,  $P = 0.022$ ), mono-n-butyl phthalate (MnBP) ( $t = -3.46$ ,  $P = 0.016$ ), and monoethylphthalate (MEP) ( $t = -2.22$ ,  $P = 0.027$ ) than male urine, and was statistically significant ( $p < 0.05$ ). The findings suggest that while the amount of PAE residue was higher in females than



**Table 1 Basic attributes of quantitative and categorical demographic variables (N=1337)**

Variable	Mean (SD)	N (%)
Gender		
Male		693(51.8)
Female		644(48.2)
Age	48.98(11.2)	
Marital status		
Married		1004(75.1)
Other		333(24.9)
Drinking habits		
Yes		105(7.9)
No (including abstinence)		1232(92.1)
Smoking habits		
Yes		311(23.2)
No (including abstinence)		1026(76.8)
Exercise habits		
Yes		568(42.5)
No		768(57.4)
Job classification		
Petrochemical manufacturing industry		129(9.6)
Construction industry		29(2.2)
Other manufacturing industries		47(3.5)
Other industries		1132(84.7)
Body mass index	24.332(3.6)	
Normal		703(52.6)
Overweight		634(47.4)
Body fat percentage	27.195(7.4)	
Normal		733(54.8)
Overweight		604(45.2)
T-score		
Normal		1113(83.2)
Abnormal (< -2.5)		224(16.8)

males among the respondents, the gender and residue amount analysis found no statistically significant difference. The correlation analysis between PAE metabolites

and regular exercise indicated that individuals who exercised regularly had lower levels of PAE metabolite MEP ( $t=0.05$ ,  $P=0.047$ ) in their urine than those who did not, with a statistically significant difference ( $p < 0.05$ ). Regular exercise and the presence of other PAE residues did not indicate a statistically significant difference.

**Risk assessment analysis of disease and PAEs**

The diseases outlined in the questionnaire samples were divided into three categories based on the systems affected. (1) Musculoskeletal system: osteoporosis, arthritis, joint pain or stiffness, neck pain, lower back and waist pain, and sciatica. (2) Cardiovascular system: coronary artery disease, cardiac arrhythmia, and cardiomyopathies. (3) Digestive and urinary systems: peptic ulcers, kidney calculi, and cholelithiasis. To assess the risk of PAE-induced diseases in these systems, the sample weight was multiplied by the tolerable daily intake (TDI). This was compared to the amount of PAE metabolites in the sample urine to determine if it exceeded the TDI [11]. We employed binary logistic regression and OR to predict the disease risk factors associated with each PAE. A Chi-square test was used to examine the correlation between PAEs and disease (Table 3).

The findings from the binary logistic regression indicate the prediction rate (PR) of the risk factor of musculoskeletal system diseases was 96.1 for osteoporosis, 95.7 for arthritis, 82.9 for joint pain or stiffness, and 94.9 for sciatica. Furthermore, we used OR to predict the risk factors of PAE disease. The results indicate that individuals with normal amounts of the PAE metabolite mono(2-ethyl-5-hydroxyhexyl) phthalate (MEHHP) (OR=0.360,  $P=0.023$ ) were 0.36 times as likely to have arthritis than those with abnormal amounts. The Chi-square analysis found significant correlations between osteoporosis and mono-(2-ethyl-5-oxohexyl) phthalate (MEOHP)

**Table 2 Relationships between PAE metabolites, gender, and regular exercise**

PAE metabolites	N=1337		Male (N=693)		Female (N=644)		Has exercise habits (N=568)		No exercise habits (N=768)	
	Normal	Excess	Mean (SD)	Mean (SD)	T-test	p-value	Mean (SD)	Mean (SD)	T-test	p-value
MEHP	227	110	17.40(29.09)	17.79(24.16)	-0.26	0.96	18.48(30.92)	16.92(23.35)	-1.05	0.137
MEOHP	151	1186	11.37(32.77)	12.05(16.2)	-0.48	0.836	12.49(22.06)	11.12(28.78)	-0.95	0.363
MEHHP	59	1279	18.79(49.66)	19.69(26.69)	-0.41	0.886	20.33(35.06)	18.40(43.73)	-0.87	0.451
MCMHP	440	897	6.94(13.09)	8.73(14.69)	-2.35	0.022*	7.93(13.00)	7.70(14.55)	-0.29	0.815
MBzP	1333	4	2.11(4.06)	2.21(3.26)	-0.46	0.698	2.33(3.92)	2.03(3.52)	0.81	0.128
MnBP	5	1332	28.21(36.68)	36.16(47.02)	-3.46	0.016*	31.67(36.26)	32.31(46.08)	0.13	0.426
MEP	27	1310	34.63(211.93)	57.1(150.57)	-2.22	0.027*	37.99(107.03)	51.01(226.29)	0.05	0.047*

Tolerable daily intake (TDI) of PAEs by the human body (mg/kg body weight/day): DEHP, 0.05; dibutyl phthalate, 0.01; di-isononyl phthalate, 0.15; butyl benzyl phthalate, 0.5; diisodecyl phthalate, 0.15 (Environmental Protection Administration Executive Yuan, Taiwan)

\*  $p < 0.05$

**Table 3 Risk assessment analysis of musculoskeletal system diseases and PAEs**

Variable	Osteoporosis			Arthritis			Joint pain or stiffness			Sciatica		
	p	OR	PR	p	OR	PR	p	OR	PR	p	OR	PR
Total			96.1			95.7			82.9			94.9
MEHP	0.144	1.592		0.447	0.916		0.134	1.276		0.355	1.218	
MEOHP	0.002*	9.123		0.223	1.583		0.037*	1.618		0.527	0.969	
MEHHP	0.363	0.737		0.023*	0.360		0.443	0.907		0.029*	0.379	
MCMHP	0.284	1.213		0.226	0.795		0.198	1.156		0.516	0.980	
MBzP	0.809	0.997		0.816	0.997		0.468	0.996				
MnBP	0.767	0.996		0.224	0.205		0.208	0.311		0.767	0.996	
MEP	0.392	0.647		0.692	1.305		0.523	1.153				

PR prediction rate, OR odds ratio, MBzP monobenzyl phthalate

\* p < 0.05

**Table 4 Risk assessment analysis of the cardiovascular system diseases and PAEs**

Variable	Coronary artery disease			Cardiac arrhythmia			Cardiomyopathies		
	p	OR	PR	p	OR	PR	p	OR	PR
Total			98.1			94.8			99.1
MEHP	0.571	0.954		0.031*	2.148		0.100	0.270	
MEOHP	0.411	2.051		0.016*	3.330		0.182	0.315	
MEHHP	0.462	0.955		0.128	3.719		0.271	0.274	
MCMHP	0.293	1.603		0.150	1.354		0.042*	0.194	
MBzP	0.950	0.997		0.786	0.997		0.979	0.997	
MnBP	0.938	0.996		0.740	0.996		0.974	0.996	
MEP	0.715	0.980		0.545	1.561		0.871	0.980	

PR prediction rate, OR odds ratio

\* p < 0.05

(OR=9.123, P=0.002); arthritis and MEHHP (OR=0.360, P=0.023); joint pain or stiffness and MEOHP (OR=1.618, P=0.037); and sciatica and MEHHP (OR=0.379, P=0.029). These findings were all statistically significant (p < 0.05) (Table 3).

Table 4 presents the analysis of various cardiovascular diseases (CVDs). The PR for coronary artery disease, cardiac arrhythmia, and myocarditis were notably high, with rates of 98.1, 94.8, and 99.1, respectively. The PAE risk factors associated with these diseases exhibited relatively high PRs. The OR results for predicting disease risk factors for the PAEs indicate that individuals with normal metabolite mono-2-ethylhexyl phthalate (MEHP) were 2.12 times more likely to exhibit cardiac arrhythmia than those with abnormal amounts. Individuals with average quantities of MEOHP were 3.34 times more likely to exhibit cardiac arrhythmia than those with abnormal quantities. In contrast, those with average amounts of MCMHP were 0.19 times more likely to exhibit cardiomyopathies than those with

abnormal quantities. The Chi-square analysis indicates a significant correlation between cardiac arrhythmia and PAE metabolites MEHP (p=0.031) and MEOHP (p=0.016). Cardiomyopathies were correlated with MCMHP (P=0.042). These findings were statistically significant (p < 0.05).

Analyzing various diseases of the digestive and urinary systems exhibited significant findings. Peptic ulcers had a PR of 85.6, kidney calculi of 89.0, and cholelithiasis of 95.3. The PAE risk factors associated with these diseases had relatively high PRs. The OR for predicting the risk factors of PAE diseases showed that individuals with average amounts of the PAE metabolite MnBP were 0.12 times more likely to exhibit peptic ulcers than those with abnormal quantities. Those with average amounts of MCMHP were 0.66 times more likely to exhibit kidney calculi than those with abnormal quantities. The Chi-square analysis showed significant correlations, with peptic ulcers correlated with the PAE metabolite MnBP (p=0.031) and kidney

calculi correlated with the PAE metabolite MCMHP ( $p=0.026$ ). The results were statistically significant ( $p < 0.05$ ).

**Artificial intelligence prediction model for disease and PAEs**

This study used a Bayesian network to analyze the main impact factors among 16 factors for disease across three systems. The Bayesian network algorithm employed predictors such as BMI and TG values related to coronary artery disease.

Figure 2 summarizes the key predictors, which included MEHP for lower back, waist, and neck pain and cardiomyopathy; MEOHP was evident for osteoporosis and cardiomyopathies; MEHHP for sciatica; MnBP was related to arthritis. Physiological data indicates that body fat was the main predictor of lower back pain and arthritis; BMI was the main predictor of sciatica. Age and bone density were the main predictors of osteoporosis, while total cholesterol was the main predictor of cardiac arrhythmia. Regarding lifestyle habits, only regular exercise was a primary predictor of lower back, waist, and neck pain. The main predictor of cardiac arrhythmia was taking hormone preparations. Regarding cardiomyopathies, the residual amounts of MEOHP (MEOHP boolean), MEHP (MEHP boolean), and regular exercise were found to be more significant, directly impacting the condition. The analysis also identified indirect impact factors. For instance, high body fat (body\_fat2) affected the residual amount of MEOHP, while low-density lipoproteins (LDLs) (LDL1) directly influenced cardiomyopathies by affecting Triglycerides (TG).

Based on the artificial neural network analysis, the overall PR exceeded 80% and was calculated as the conditional probability of 18 factors for each disease. For example, with a PR of 75.8, the main predictors of lower back and waist pain included MEP, body fat, and smoking. The main predictors of sciatica, with a PR of 94.7, included MEOHP, age, and drinking. Moreover, the main predictors of neck pain, with a PR of 71.1, were MEHP and regular exercise, while the main predictors of arthritis, with a PR of 94.9, included MEHHP, MCMHP, and body fat. Osteoporosis, with a PR of 94.7, had the following predictors: MEHP, MEP, age, bone density, BMI, and smoking.

**Comparing prediction model AUCs**

We used the Bayesian and artificial neural networks and XG boots to compare the predictors of musculoskeletal system disease risk factors. Their prediction performances were assessed based on the areas under the curves (AUCs), and no significant differences were

**Table 5 The area under the curve (AUC) for three models**

Test result variable	Area	Standard error	Asymptotic significance	95% confidence interval
Bayes	0.827	0.054	0.000	0.720–0.933
XG boots	0.634	0.090	0.138	0.457–0.810
Neural network	0.502	0.052	0.970	0.399–0.605

indicated. The AUC of the artificial neural network’s prediction performance was 0.502. The Bayesian network prediction performance was 0.827. After using XG boots to test the data in the dataset, we found that the overall prediction recall rate was 0.634, the precision was 0.577, and the F1-score was above 0.704. All AUCs were  $> 0.5$ . This indicates that the classifiers outperformed random guessing. However, the Bayesian network slightly outperformed the others (Table 5) (Fig. 3).

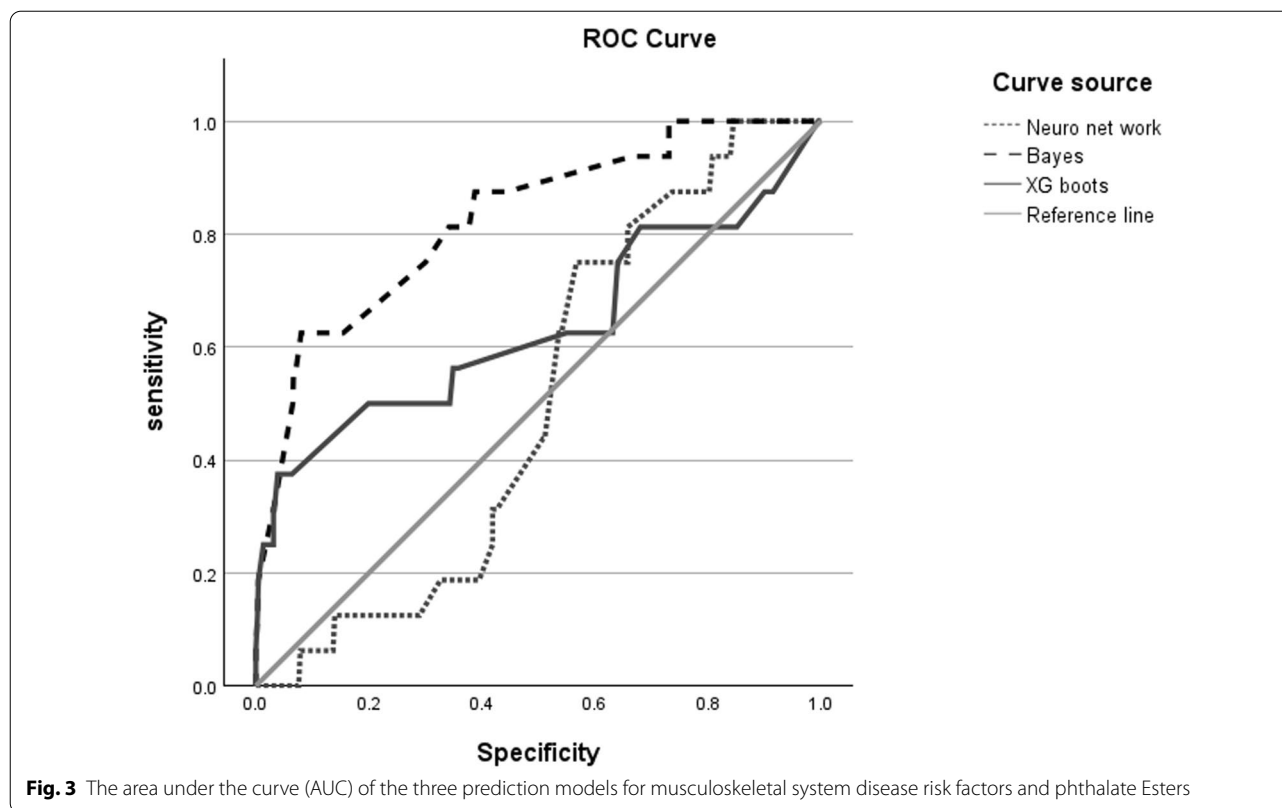
**Discussion**

**Difference analysis and discussion of gender, regular exercise, and PAE metabolites**

In a retrospective human biomonitoring study, Wittassek et al. [19] examined urine samples from the German Environmental Specimen Bank for Human Tissues from 1988 to 2003. They discovered a significant increase in females’ daily intake of dibutyl phthalate (DnBP, DiBP), aligning with the present study’s findings [20]. We observed higher amounts of metabolites in females’ urine than in males’ urine, including MCMHP (a metabolite of DEHP), MnBP (a metabolite of DnBP), and MEP (a metabolite of DEP). PAE DnBP is commonly used in products such as food packaging, latex, adhesives, and product solvents, including cosmetics and lotions, which are popular among females. This could explain why females exhibited higher metabolite levels than males.

Buser et al. [2] reported age and gender differences in the correlations between urinary phthalate metabolite concentrations and body weight, aligning with the present study’s findings. While this study found gender differences in the PAE metabolite content of urine, it did not observe differences based on age. Analyzing regular exercise and PAE metabolites, we discovered lower levels of PAE metabolite MEP in respondents who exercised regularly than in those who did not. However, no statistical difference was observed in the analysis of regular exercise and other PAE residues. Meanwhile, no other study on associated topics was found. Primarily, the residue of PAE metabolites is influenced by various factors, such as age, gender, body fat percentage, metabolic organ





functions, exposure dose, and time of sample collection, all of which can lead to data errors [21, 22]. Therefore, this study analyzed the results from data samples from the natural environment.

#### Differences in risk factor predictions

In examining CVD predictors, Wen et al. [20] reported that DEHP and its metabolites significantly affect CVD-related processes and factors, including cardiac developmental toxicity, cardiac injury and apoptosis, cardiac arrhythmia, cardiometabolic disorders, the structural injury of blood vessels, atherosclerosis, coronary heart disease, and hypertension [23]. These findings align with the present study. Using Bayesian network modeling, the present study identified MEHP, MEOHP, and LDL as critical predictors of cardiomyopathy. Therefore, PAE metabolites can be considered as coronary heart disease predictors based on past and current findings.

Dong et al. [24] used logistic regression analysis to determine the correlation between urinary phthalate metabolites, obesity, and concentric obesity. They found that the MMP, MEHHP, and MECPP urine contents were associated with increased proportions of concentric obesity [25], aligning with the results of this study. The Bayesian network analysis in the present study discovered that

MEOHP and MEHP residue directly affected cardiomyopathies. Indirect factors such as high body fat (body\_fat2) were found to influence the MEOHP residual amount, indirectly causing cardiomyopathies. Based on past studies and the present findings, PAE metabolites can serve as predictors of coronary heart disease alongside common factors such as age and cholesterol.

Literature on musculoskeletal and joint disease predictors indicates that specific phthalate biomarkers are associated with decreased bone density in the main hip joint and femoral neck [26]. Animal studies also suggest that DEHP negatively affects bone metabolism in ovariectomized rats, indicating its negative impact on the bone metabolism of postmenopausal females [27, 28]. In an experimental study, Carwile et al. [5] used linear regression to establish the correlation between biomarkers of various polyfluoroalkyl substances (PFASs), phthalate, and areal bone mineral density (aBMD). Bayesian network regression has also been used to examine the correlation between PFAS/phthalate integral biomarker mixture and aBMD. Findings indicate that phthalates could be correlated with decreased bone mineral density in adolescent males. These findings suggest that PAE metabolites are correlated with bone density, aligning with the results of the present study. Furthermore, decreased estrogen secretion in postmenopausal females

indirectly affects bone density. Using a Bayesian network and an artificial neural network to develop models [29], we found that MEOHP and bone density directly affected osteoporosis (based on the Bayesian network). Moreover, indirect factors such as menopause were found to influence bone density, indirectly causing osteoporosis. The artificial neural network model suggests that MEHP, MEP, age, bone density, BMI, and smoking are the main predictors of osteoporosis. Therefore, based on prior research and the present study, a correlation between PAE metabolites and bone density can predict musculoskeletal and joint diseases [30].

Guo conducted a cross-sectional study using data from the 2011–2018 National Health and Nutrition Examination Survey (NHANES) to examine the correlation between OPE exposure and bone mineral density (BMD) in U.S. adults. They found a negative correlation between urinary OPE metabolite concentration and BMD. Their findings also suggest that men are more vulnerable than women [31]. A study on female osteoporosis found evidence to support that endocrine-disrupting chemicals (EDCs) commonly used in plastics and personal care products may affect vitamin D absorption, indirectly affecting bone density [32].

#### **Suggestions for improved nursing assessment**

The present study found that, in addition to age, bone density, BMI, body fat, smoking, drinking, and regular exercise, PAE metabolites serve as predictors of disease among individuals with musculoskeletal disorders. Nursing workloads can be reduced by optimizing existing nursing assessment systems, while prediction accuracy for individual nursing diagnoses and disease risks can be improved [29]. The predictors identified in this study can be integrated into nursing assessment content during hospital admissions, along with the patient's medical history, clinical symptoms, examination reports, and dietary information. Most current hospital nursing assessment items are general medical history inquiries. We recommend that hospitals include predictors identified in this study in nursing assessments to facilitate early prediction of patients' risk of exposure and excessive accumulation of plasticizers. For instance, the basic information diagram includes smoking and drinking history. It can also include exercise habits, frequency, and duration. Regarding medical assessments, medication options, and types can be added, and hormone preparation can be included. Regarding gastrointestinal system assessments, body fat percentage should be assessed in addition to height, weight, and BMI. Items such as dietary type, eating-out habits, and frequency should be added to the dietary section of these assessments.

#### **Conclusions**

This study used machine learning algorithms to examine critical factors related to PAEs, physiological data, and lifestyle habits for various diseases. Using the conditional probability of 18 factors for each disease, we found that PAEs significantly affected osteoarthritis with a lesser impact on heart disease. Gender differences were observed in PAE metabolite residue, with females exhibiting higher PAEs in MCMHP, MnBP, and MEP residues than males, indicating a statistically significant difference. The analysis regarding regular exercise and PAE metabolites showed that individuals who exercised regularly had lower levels of MEP than those who did not, indicating a statistically significant difference.

#### **Limitations**

The study limitations are as follows. First, this study relied on data from the Taiwan Biobank, which only included 1337 complete cases for individuals aged 30 to 70. Future studies should consider the retrospective nature of the analysis, potential confounding factors not addressed, and the generalizability of findings to populations outside of Taiwan. Moreover, current routine medical tests in Taiwan do not include the examination of environmental hormones. This is due to factors such as testing fees not covered by health insurance and a lack of public awareness regarding the potential health effects of environmental hormones. Providing these study results to the government can increase awareness, potentially leading to policy changes and public health initiatives to promote testing for environmental hormones. Furthermore, combining data from various sources, such as the health insurance database or specialized testing units, can augment the dataset's comprehensiveness and the prediction model's accuracy.

#### **Acknowledgements**

We thank the staff of the Teaching and Research Center, Cheng Hsin General Hospital, for their assistance in the research process.

#### **Funding**

Financial support for this study was provided by a grant from Cheng Hsin General Hospital (Grant No. CHGH 110D010-05). The funding agreement ensured the authors' independence in designing the study, interpreting the data, and writing and publishing the report.

#### **Data availability**

The datasets analyzed during the current study are not publicly available due to IRB restrictions, but are available from the corresponding author on reasonable request.

#### **Declarations**

#### **Conflict of interest**

The authors have not disclosed any competing interests.

### Ethics approval

This study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Research Ethics Committee of the Taiwan Biobank approved the study before it was conducted (Approval Number: TWBR11007-06). All procedures performed in studies involving human participants were conducted following the ethical standards of the Institutional and National Research Committee per the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards. This study was approved by the Research Ethics Committee of the Taiwan Biobank (approval number: TWBR11007-06). Since the data of this study were sourced from a database, informed consent was waived by the Institutional Review Board of Cheng Hsin General Hospital, Taiwan (IRB number: (862)110-08) upon approving the Informed Consent Form Waiver. Data from the Taiwan Biobank are voluntarily provided by subjects participating in examinations across Taiwan. A researcher cannot apply to the Taiwan Biobank until the institution's IRB approves a study. Delinked data where personal information has been removed can be obtained after a study is reviewed and approved and payment is made. All data are destroyed within two years after a study is published.

### Informed consent

As the data of this study were sourced from a database, the informed consent was waived by the Institutional Review Board of Cheng Hsin General Hospital, Taiwan (IRB number: (862)110-08) upon approving the Informed Consent Form Waiver.

### Author details

<sup>1</sup>Department of Nursing, Cheng Hsin General Hospital, Taipei, Taiwan. <sup>2</sup>Department of Gastroenterologist, Tainan Municipal Hospital (Managed by Show Chwan Medical Care Corporation), Tainan, Taiwan. <sup>3</sup>Department of Research, Taiwan Academy of Ecological Hazard & Health Management, Taipei, Taiwan. <sup>4</sup>Nursing Reserach center & School of Nursing, Taipei Medical University Hospital & Taipei Medical University, Taipei, Taiwan. <sup>5</sup>School of Nursing, National Taipei University of Nursing and Health Sciences, No. 365, Ming-Te Road, Peitou District, Taipei 112, Taiwan.

Received: 11 April 2024 Accepted: 10 December 2024

Published online: 29 December 2024

### References

- Bopp SK, Barouki R, Brack W, Dalla Costa S, Dorne JCM, Drakvik PE, Faust M, Karjalainen TK, Kephelopoulous S, van Klaveren J, Kolossa-Gehring M, Kortenkamp A, Lebret E, Lettieri T, Nørager S, Rüegg J, Tarazona JV, Trier X, van de Water B, van Gils J, Bergman Å. Current EU research activities on combined exposure to multiple chemicals. *Environ Int.* 2018;120:544–62. <https://doi.org/10.1007/s00339-007-4137-z>.
- Buser MC, Murray HE, Scinicariello F. Age and sex differences in childhood and adulthood obesity association with phthalates: analyses of NHANES 2007–2010. *Int J Hyg Environ Health.* 2014;217:687–94. <https://doi.org/10.1016/j.ijheh.2014.02.005>.
- Brehm E, Flaws JA. Transgenerational effects of endocrine-disrupting chemicals on male and female reproduction. *Endocrinology.* 2019;160:1421–35. <https://doi.org/10.1210/en.2019-00034>.
- Botton J, Kadawathagedara M, de Lauzon-Guillain B. Endocrine disrupting chemicals and growth of children. *Ann Endocrinol.* 2017;78:108–11. <https://doi.org/10.1016/j.ando.2017.04.009>.
- Carwile JL, Seshasayee SM, Ahrens KA, Hauser R, Driban JB, Rosen CJ, Gordon CM, Fleisch AF. Serum PFAS and urinary phthalate biomarker concentrations and bone mineral density in 12–19 year olds: 2011–2016 NHANES. *J Clin Endocrinol Metab.* 2022;107:e3343–52. <https://doi.org/10.1210/clinem/dgac228>.
- Encarnaçao T, Pais AA, Campos MG, Burrows HD. Endocrine disrupting chemicals: impact on human health, wildlife and the environment. *Sci Prog.* 2019;102:3–42. <https://doi.org/10.1177/0036850419826802>.
- Gharani P, Suffoletto B, Chung T, Karimi HA. An artificial neural network for movement pattern analysis to estimate blood alcohol content level. *Sensors.* 2017;17:2897. <https://doi.org/10.3390/s17122897>.
- Guo Y, Kannan K. A survey of phthalates and parabens in personal care products from the United States and its implications for human exposure. *Environ Sci Technol.* 2013;47:14442–9.
- Li T, Yu Y, Sun Z, Duan J. A comprehensive understanding of ambient particulate matter and its components on the adverse health effects based from epidemiological and laboratory evidence. *Part Fibre Toxicol.* 2022;19:67. <https://doi.org/10.1186/s12989-022-00507-5>.
- Rettenmeier A, Drexler H, Hartwig A. The MAK-collection for occupational health and safety. German: Wiley; 2022.
- Liu Y, Li X, Pu Q, Fu R, Wang Z, Li Y, Li X. Innovative screening for functional improved aromatic amine derivatives: toxicokinetics, free radical oxidation pathway and carcinogenic adverse outcome pathway. *J Hazard Mater.* 2023;454: 131541. <https://doi.org/10.1016/j.jhazmat.2023.131541>.
- Wan MLY, Co VA, El-Nezami H. Endocrine disrupting chemicals and breast cancer: a systematic review of epidemiological studies. *Crit Rev Food Sci Nutr.* 2022;62:6549–76.
- Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, Moons KGM, Collins G, van Smeden M. Calculating the sample size required for developing a clinical prediction model. *BMJ.* 2020;368: m441. <https://doi.org/10.1136/bmj.m441>.
- Shipe ME, Deppen SA, Farjah F, Grogan EL. Developing prediction models for clinical use using logistic regression: an overview. *J Thorac Dis.* 2019;11:S574–84.
- LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521:436–44. <https://doi.org/10.1038/nature14539>.
- Chien PL, Liu CF, Huang HT, Jou HJ, Chen SM, Young TG, Wang YF, Liao PH. Application of artificial intelligence in the establishment of an association model between metabolic syndrome, TCM constitution, and the guidance of medicated diet care. *Evid-Based Complement Altern Med.* 2021;2021:5530717. <https://doi.org/10.1155/2021/5530717>.
- Chu W, Ho C-S, Liao P-H. Comparison of different predicting models to assist the diagnosis of spinal lesions. *Inf Health Soc Care.* 2022;47:92–102. <https://doi.org/10.1080/17538157.2021.1939355>.
- Chen, T., Guestrin, C.: (2016) XGBoost: a scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, pp. 785–794. <https://doi.org/10.1145/2939672.2939785>
- Wittassek M, Wiesmüller GA, Koch HM, Eckard R, Döbler L, Müller J, Angerer J, Schlüter C. Internal phthalate exposure over the last two decades—a retrospective human biomonitoring study. *Int J Hyg Environ Health.* 2007;210:319–33. <https://doi.org/10.1016/j.ijheh.2007.01.037>.
- Wen ZJ, Wang ZY, Zhang YF. Adverse cardiovascular effects and potential molecular mechanisms of DEHP and its metabolites—a review. *Sci Total Environ.* 2022;847: 157443. <https://doi.org/10.1016/j.scitotenv.2022.157443>.
- Tassinari R, Tait S, Busani L, Martinelli A, Valeri M, Gastaldelli A, Deodati A, La Rocca C, Maranghi F. Toxicological assessment of oral co-exposure to bisphenol A (BPA) and bis(2-ethylhexyl) phthalate (DEHP) in juvenile rats at environmentally relevant dose levels: evaluation of the synergic, additive or antagonistic effects. *Int J Environ Res Public Health.* 2021;18:4584. <https://doi.org/10.3390/ijerph18094584>.
- Rettenmeier A, Drexler H, Hartwig A. Di(2-ethylhexyl) phthalate (DEHP) [BAT value documentation, 2018]. MAK-Collect Occup Health Saf. 2022;4:906–20. <https://doi.org/10.1002/3527600418.bb11781e2319>.
- Su T-C, Hwang J-J, Sun C-W, Wang SL. Urinary phthalate metabolites, coronary heart disease, and atherothrombotic markers. *Ecotoxicol Environ Saf.* 2019;173:37–44. <https://doi.org/10.1016/j.ecoenv.2019.02.021>.
- Dong R, Zhou T, Chen J, Zhang M, Zhang H, Wu M, Li S, Zhang L, Chen B. Gender- and age-specific relationships between phthalate exposures and obesity in Shanghai adults. *Arch Environ Contam Toxicol.* 2017;73:431–41. <https://doi.org/10.1007/s00244-017-0441-6>.
- Mohanto NC, Ito Y, Kato S, Kamijima M. Life-time environmental chemical exposure and obesity: review of epidemiological studies using human biomonitoring methods. *Front Endocrinol.* 2021;12: 778737. <https://doi.org/10.3389/fendo.2021.778737>.
- Reeves KW, Vieyra G, Grimes NP, Meliker J, Jackson RD, Wactawski-Wende J, Wallace R, Zoeller RT, Bigelow C, Hankinson SE, Manson JE, Cauley JA, Calafat AM. Urinary phthalate biomarkers and bone mineral density in postmenopausal women. *J Clin Endocrinol Metab.* 2021;106:e2567–79.
- Choi JI, Cho HH. Effects of di(2-ethylhexyl) phthalate on bone metabolism in ovariectomized mice. *J Bone Metab.* 2019;26:169–77.

28. Lai CC, Liu FL, Tsai CY, Wang SL, Chang DM. Di-(2-ethylhexyl) phthalate exposure links to inflammation and low bone mass in premenopausal and postmenopausal females: evidence from ovariectomized mice and humans. *Int J Rheum Dis*. 2022;25:926–36. <https://doi.org/10.1111/1756-185X.14386>.
29. Liao P-H, Tsuei Y-C, Chu W. Application of machine learning in developing decision-making support models for decompressed vertebroplasty. *Healthcare*. 2022;10:214. <https://doi.org/10.3390/healthcare10020214>.
30. Chen Y-T, Chiu Y-C, Teng M-L, Liao P-H. The effect of medical material management system APP on nurse workload and stress. *BMC Nurs*. 2022;21:243. <https://doi.org/10.1186/s12912-021-00767-0>.
31. Guo JY, Wang SN, Zhang ZL, Luan M. Associations between organophosphate esters and bone mineral density in adults in the United States: 2011–2018 NHANES. *Ecotoxicol Environ Saf*. 2024;278: 116414. <https://doi.org/10.1016/j.ecoenv.2024.116414>.
32. Brennan E, Butler AE, Nandakumar M, Thompson K, Sathyapalan T, Atkin SL. Relationship between endocrine disrupting chemicals (phthalate metabolites, triclosan and bisphenols) and vitamin D in female subjects: an exploratory pilot study. *Chemosphere*. 2024;349: 140894. <https://doi.org/10.1016/j.chemosphere.2023.140894>.