SCHD
Structural & Congenital Heart Disease

Tech Science Press

http://doi.org/10.32604/schd.2026.085942

ARTICLE

Unsafe Sanitation and the Global Incidence of Congenital Heart Disease: A Spatial Correlation Analysis

Yi Shen[1,2,3,#], Zeye Liu[4,#], Jing Xie[5,#], Xuanqi An[6,7,8,#], Zeyu Jing[1,2,3], Wenchuan Liao[1,2,3], Yifan Zhu[1,2,3], Chenyu Jiang[1,2,3], Xingliang Zhou[1,2,3], Xu Huang[1,2,3], Tianyu Liu[1,2,3], Jian Liu[1,2,3], Yuxi Ji[1,2,3], Yi Yan[1,3], Bei Feng[1,3], Yiwei Liu[1,2,3], Yi Shi[4,*], Yanjun Sun[2,*] and Hao Zhang[1,2,3,*]

[1]Heart Center and Shanghai Institute of Pediatric Congenital Heart Disease, Shanghai Children's Medical Center, Nationa Children's Medical Center, Shanghai Jiaotong University School of Medicine, Shanghai, China
[2]Department of Cardiothoracic Surgery, Shanghai Children's Medical Center, National Children's Medical Center, Shanghai Jiaotong University School of Medicine, Shanghai, China
[3]Shanghai Clinical Research Center for Rare Pediatric Diseases, Shanghai Children's Medical Center, National Children's Medical Center, Shanghai Jiaotong University School of Medicine, Shanghai, China
[4]Department of Cardiac Surgery, Peking University People's Hospital, Peking University, Xicheng District, Beijing, China
[5]Department of Pharmacy, Zhongda Hospital, School of Medicine, Southeast University, Nanjing, China
[6]Emergency Department, National Clinical Research Center of Cardiovascular Diseases, Fuwai Hospital, National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China
[7]State Key Laboratory of Cardiovascular Disease, Fuwai Hospital, National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China
[8]National Clinical Research Center for Cardiovascular Diseases, Fuwai Hospital, Chinese Academy of Medical Sciences, Beijing, China
*Corresponding Authors: Yi Shi. Email: shiyi1088@126.com; Yanjun Sun. Email: sunyanjuan@scmc.com.cn; Hao Zhang. Email: drzhanghao@126.com
[#]These authors contributed equally to this work and are joint first authors

Received: 21 May 2026; Accepted: 27 July 2026; Published: 31 July 2026

ABSTRACT: Background: Congenital heart disease (CHD) is the most common congenital anomaly worldwide, yet the contribution of environmental factors to its global geographic variation remains incompletely understood. We aimed to systematically identify environmental factors associated with CHD incidence using an integrated framework combining machine learning and spatial epidemiology. **Methods:** Country-level data were obtained from the Global Burden of Disease (GBD) 2021 study. Boruta algorithm-based feature selection and random forest SHAP value ranking were applied to identify environmental factors associated with CHD incidence. Negative binomial regression was used to evaluate the associations between selected variables and CHD incidence. Spatial clustering was assessed using Global Moran's I and Local Indicators of Spatial Association (LISA). The population attributable fraction (PAF) associated with unsafe sanitation was further estimated. **Results:** Boruta identified 28 candidate environmental variables associated with CHD incidence, among which unsafe sanitation consistently ranked as the most influential factor based on SHAP analysis. Negative binomial regression demonstrated a significant association between unsafe sanitation and CHD incidence. Spatial analyses revealed marked geographic concordance between regions with high CHD incidence and high unsafe sanitation exposure, particularly in Sub-Saharan Africa. The estimated PAF suggested that approximately 33% of the global CHD burden could theoretically be attributable to unsafe sanitation under the assumptions of the analytical model. **Conclusions:** Unsafe sanitation was consistently identified as the environmental factor most strongly associated with global CHD incidence across multiple analytical approaches. These findings provide ecological evidence supporting a potential relationship between sanitation conditions and

the geographic distribution of CHD. However, further individual-level epidemiological and mechanistic studies are required to validate these findings and clarify the underlying biological mechanisms.

KEYWORDS: Congenital heart disease; unsafe sanitation; machine learning; spatial epidemiology; spatial autocorrelation; boruta algorithm; population attributable fraction

1 Introduction

Congenital heart disease (CHD) is the most common birth defect worldwide, and it is one of the main causes of infant mortality and long-term disability [1,2]. It is estimated that each year over 1.3 million newborns worldwide have CHD, and tens of thousands of cases require medical intervention in the early stages of life, imposing a heavy burden on families and society [1,3]. Although the diagnosis rate and survival rate of CHD have significantly improved with advancements in medical imaging, surgical techniques, and perinatal management, its incidence has remained high globally and varied widely across regions over the past few decades, especially in low- and middle-income countries [2]. The prevalence and mortality rates of CHD in low socio-demographic index (SDI) areas are significantly higher than those in high SDI areas [4,5]. This trend suggests that, in addition to genetic susceptibility factors, geographical differences in environmental and socio-economic related factors may play an important role in the occurrence of CHD.

The etiology of CHD is complex. Previous mechanistic studies have discovered that, in addition to genetic mutations, environmental hazards play an important role in pathogenesis. Only about 15% of CHD cases can be attributed to determined genetic causes, while at least 10% of CHD cases are directly related to environmental factors [6,7]. Epidemiological studies have extensively explored environmental exposure factors that affect the risk of CHD, such as air pollution, heavy metals, maternal infections, drug use, and unhealthy lifestyles [8–12]. However, most of these studies have focused on specific environmental toxins or macroscopic socioeconomic indicators, such as per capita GDP or urbanization rate, while largely ignoring a more fundamental and highly modifiable public health factor: basic sanitation.

According to the World Health Organization (WHO), unsafe sanitation includes open defecation, the lack of private toilets or latrines, and the lack of facilities to safely isolate human excrement from the environment [13]. Such conditions not only lead to a high incidence of intestinal infectious diseases but may also indirectly affect pregnancy outcomes and fetal development. Cardiac morphogenesis primarily occurs during weeks 3–8 of human gestation, encompassing cardiac looping, chamber formation, septation, and outflow tract development. During this critical developmental window, the embryonic heart is particularly susceptible to maternal environmental perturbations, including infection, systemic inflammation, nutritional deficiencies, and toxic environmental exposures [14,15]. Unsafe sanitation may contribute to these adverse maternal conditions through increased exposure to infectious pathogens, impaired nutritional status, altered maternal microbiota, and environmental contaminants. For example, poor sanitation is strongly associated with increased maternal exposure to infectious agents and inflammatory conditions during pregnancy, which are established risk factors for CHD [16,17]. In fact, there are still over 1.5 billion people worldwide who lack access to safe sanitation facilities, and a considerable portion of them are women of childbearing age [13]. Currently, most epidemiological models of birth defects only consider traditional environmental or behavioral exposures, such as smoking, drinking, and air pollution exposure, while neglecting this fundamental but widespread inequality factor of hygiene conditions.

The aim of this study is to explore the relationship between exposure to unsafe sanitation facilities during reproductive age and the incidence of CHD from the perspective of spatial distribution, using data from the Global Burden of Disease (GBD), Injuries, and Risk Factors Study 2021 database.

2 Methods

2.1 Overview

Comprehensive data from the GBD 2021 database were utilized in this study to examine the spatial distribution of CHD and its related risk factors. The analysis covered the incidence of CHD across all age groups in 204 countries and regions, as well as the summary exposure values (SEV) for risk factors among women of childbearing age (15–49 years). This manuscript was produced as part of the GBD Collaborator Network and by GBD protocols (Contact ID: 0034o00001nHH4NAAW) and followed the GATHER guidelines [18].

Comprehensive statistical analysis and machine learning models were employed, including global Moran's I analysis, local indicators of spatial association (LISA) analysis, the Boruta algorithm based on random forests, Shapley additive explanations (SHAP) analysis, and negative binomial regression, to identify key risk factors, clarify their spatial distribution relationship with the occurrence of CHD, and quantify their impact on disease burden.

2.2 Data Extraction

The incidence data for CHD and the exposure data for risk factors were all obtained from the GBD database. The detailed information on the data search, extraction, and screening criteria of GBD 2021 has been described elsewhere [19]. In brief, the GBD database collects data on 371 diseases and injuries, as well as 88 risk factors worldwide, covering information across different age groups, genders, and regions, including assessment of disease burden, disease trends, and risk factors [19]. Congenital heart malformations include all live births with existing structural abnormalities of the heart and great vessels at birth, including but not limited to abnormalities in the structure of the heart (valvular abnormalities, ventricular or atrial septal defect, single ventricle heart defects, etc.), abnormalities in the structure of major blood vessels (malformations of great vessels, patent ductus arteriosus, etc.), and other complex congenital heart defects. In this study, we extracted the incidence of CHD for all age groups in 204 countries and regions in 2021, as well as the exposure rates of all 58 most detailed risk factors for women of childbearing age (aged 15 to 49), including environmental, occupational, and lifestyle factors.

2.3 Risk Factor Selection

Only risk factors with standardized Summary Exposure Values (SEVs) for women aged 15–49 years available in GBD 2021 were eligible for machine-learning screening. Several important maternal-level risk factors, including maternal diabetes, obesity, prenatal care utilization, and genetic susceptibility, are not currently represented by globally standardized SEVs and therefore could not be incorporated into the present analytical framework. To more comprehensively identify risk factors for the onset of CHD, we employed a three-step method that independently screened all 58 most detailed risk factors. Firstly, we used the Boruta algorithm based on random forest to calculate the importance scores for each risk factor. To ensure the robustness and reproducibility of the feature selection process, a fixed random seed was specified, a significance threshold of $p = 0.01$ with multiple-testing correction was applied, and the maximum number of iterations was set to 200. Variables classified as tentative were subsequently evaluated using the

TentativeRoughFix procedure, and only variables confirmed as important were retained for downstream analyses.

Then, 28 important risk factors with significantly higher importance scores than those of randomly generated shadow features were selected from the 58 most detailed risk factors. Subsequently, for the remaining 28 risk factors, the SHAP method, a machine learning algorithm, was applied to assess the influence of each risk factor, thereby further screening those potentially associated with CHD.

Finally, we used negative binomial regression to estimate the incidence rate ratio (IRR) for each risk factor and identify the one most strongly associated with the onset of CHD. The dependent variable in the negative binomial regression model was the number of incident cases (*val.num*). To account for differences in population size across countries and regions, the natural logarithm of the population size [log(pop)] was included as an offset term, allowing the model to estimate incidence rate ratios (IRRs). The covariates included unsafe sanitation, household air pollution from solid fuels, no access to handwashing facilities, unsafe water source, lead exposure, occupational exposure to particulate matter, gases and fumes, high temperature, and the 2021 SDI. All covariates were standardized prior to model fitting. The Benjamini-Hochberg (BH) procedure was applied to control the false discovery rate (FDR) for the statistical tests involving multiple variables.

2.4 Definition of the Spatial Distribution Characteristics of Congenital Heart Anomalies and Unsafe Sanitation

In order to explore the spatial distribution and spatial clustering patterns of CHD and the identified risk factors, we first divided the CHD incidence level and the exposure levels of the identified risk factors into 4 levels: low (<25 percentiles), lower-middle (25–50 percentiles), upper-middle (50–75 percentiles), and high (>75 percentiles). Subsequently, we used global Moran's I analysis to analyze the overall spatial autocorrelation of CHD and the identified risk factors, with the calculation formula as follows:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \cdot \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \overline{x})\left(x_j - \overline{x}\right)}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

here, n is the sample size, x_i is the attribute value of region i, $\overline{x}$ is the average value, and w_{ij} is the spatial weight used to describe the adjacency relationship between i and j. Before calculation, the normality of the data was tested. If the data did not meet the normal distribution, the calculation of the p-value was conducted using the Monte Carlo permutation test. Then, we further used LISA analysis to calculate the individual autocorrelation index for each spatial unit and divide them into four key regions: hotspots with high values surrounded by high values (High-High), cold spots with low values surrounded by low values (Low-Low), and spatial outliers where high values are surrounded by low values or *vice versa* (High-Low or Low-High).

2.5 Disease Attribution

For the identified risk factors, we calculated the PAF using a continuous exposure-response framework. Specifically, the dose-response relationship between each risk factor and CHD incidence was estimated using a negative binomial regression model. All exposure variables were standardized before model fitting, and the regression coefficient (β) was used to construct a continuous risk function:

$$RR(x) = e^{\beta x},$$

where x represents the standardized exposure level.

Because no established theoretical minimum risk exposure level (TMREL) is currently available for these environmental risk factors, we defined the counterfactual exposure distribution empirically. Specifically, individuals with exposure levels below the population median were considered the low-exposure reference group, and the empirical exposure density of this group was used to estimate $P^*(x)$. Therefore, the formula for calculating the *PAF* of the jth ($j = 1,\ 2$) disease in the kth ($k = 1,\ 2,\ 3$) region for the ith risk factor is as follows:

$$PAF = \frac{\int_{min}^{max} RR_{ij}(x)P_{ik}(x)dx - \int_{min}^{max} RR_{ij}(x)P_i{}^*(x)dx}{\int_{min}^{max} RR_{ij}(x)P_{ik}(x)dx},$$

among them, x represents the risk factor exposure rate, min and max denote the lowest and highest exposure levels for each risk factor, and $RR_{ij}(x)$ is calculated by the negative binomial regression for the IRR value of the ith risk factor at the exposure rate x for the jth disease incidence. $P_{ik}(x)$ is the exposure rate distribution of the ith risk factor in the kth region, and $P_i{}^*(x)$ is the theoretical distribution of the minimum risk exposure level of the ith risk factor, which is estimated based on the exposure level distribution of countries and territories with exposure levels lower than the median of the global exposure rate.

3 Results

3.1 Identification of Key Environmental Associated Factors of CHD

In order to identify the main risk factors for women of childbearing age, we first employed the Boruta algorithm to screen the 58 most detailed risk factors for women aged 15–49 years old. Based on the importance results, a total of 28 significant risk factors were identified out of the 58 most detailed risk factors for women of childbearing age (Fig. 1). The importance scores of these features were significantly higher than those of randomly generated shadow features, indicating that they have stable and significant predictive contributions in the model. Among them, unsafe sanitation, no access to handwashing facilities, household air pollution from solid fuels, and unsafe water sources had the highest importance scores, suggesting that unsatisfactory basic hygiene conditions are the main risk sources for women aged 15–49. Secondly, several occupational and environmental exposure factors also exhibit high importance, such as lead exposure, occupational exposure to asbestos, occupational particulate matter gases and fumes, and occupational exposure to polycyclic aromatic hydrocarbons, etc. To further evaluate model performance, a random forest model was constructed using the Boruta-selected variables and assessed by 10-fold cross-validation. Model performance was evaluated using the root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2). The cross-validation results demonstrated good predictive performance, with an RMSE of 9.589, an MAE of 6.872, and an R^2 of 0.841, suggesting that the model achieved satisfactory predictive accuracy while reducing the risk of overfitting.

The algorithm identifies 28 variables with statistically significant predictive importance for congenital heart disease (CHD) incidence, shown by bars in green.

Next, to reveal the specific contribution directions and intensities of each risk factor to disease risk prediction, we used the SHAP method to conduct a further analysis on the 28 significant risk factors identified. The analysis results showed that the contribution directions and intensities of different features in the model's predictions differed significantly (Fig. 2). We selected seven risk factors with higher SHAP values, which were concentrated in the positive region and associated with increased disease risk. The risk factors include: unsafe sanitation, household air pollution from solid fuels, no access to handwashing

facilities, unsafe water source, lead exposure, occupational particulate matter, gases, and fumes, and high temperature. Among them, the SHAP values of unsafe sanitation, no access to handwashing facilities, and unsafe water source were the highest among all variables, suggesting that the lack of sanitation facilities and inadequate living environmental conditions are the strongest risk drivers. Air pollution, lead exposure, and occupational exposure also showed positive risk trends and played an important role in risk prediction. These results indicated that the contributions of the macro-environment and lifestyle factors to the risk were much greater than those of individual behaviors or noise factors. The lack of sanitation facilities and inadequate hygiene conditions may occupy the most prominent position in the occurrence of diseases. Because the GBD database primarily includes established or suspected risk factors, variables exhibiting an inverse association with CHD incidence in regression analyses may reflect residual confounding, data heterogeneity, or limitations inherent to ecological analyses rather than genuine protective effects. Therefore, to facilitate interpretation of independent associations among plausible risk factors, variables demonstrating consistent positive contributions in the SHAP analysis and biological plausibility as environmental risk factors were prioritized for subsequent multivariable regression analyses.

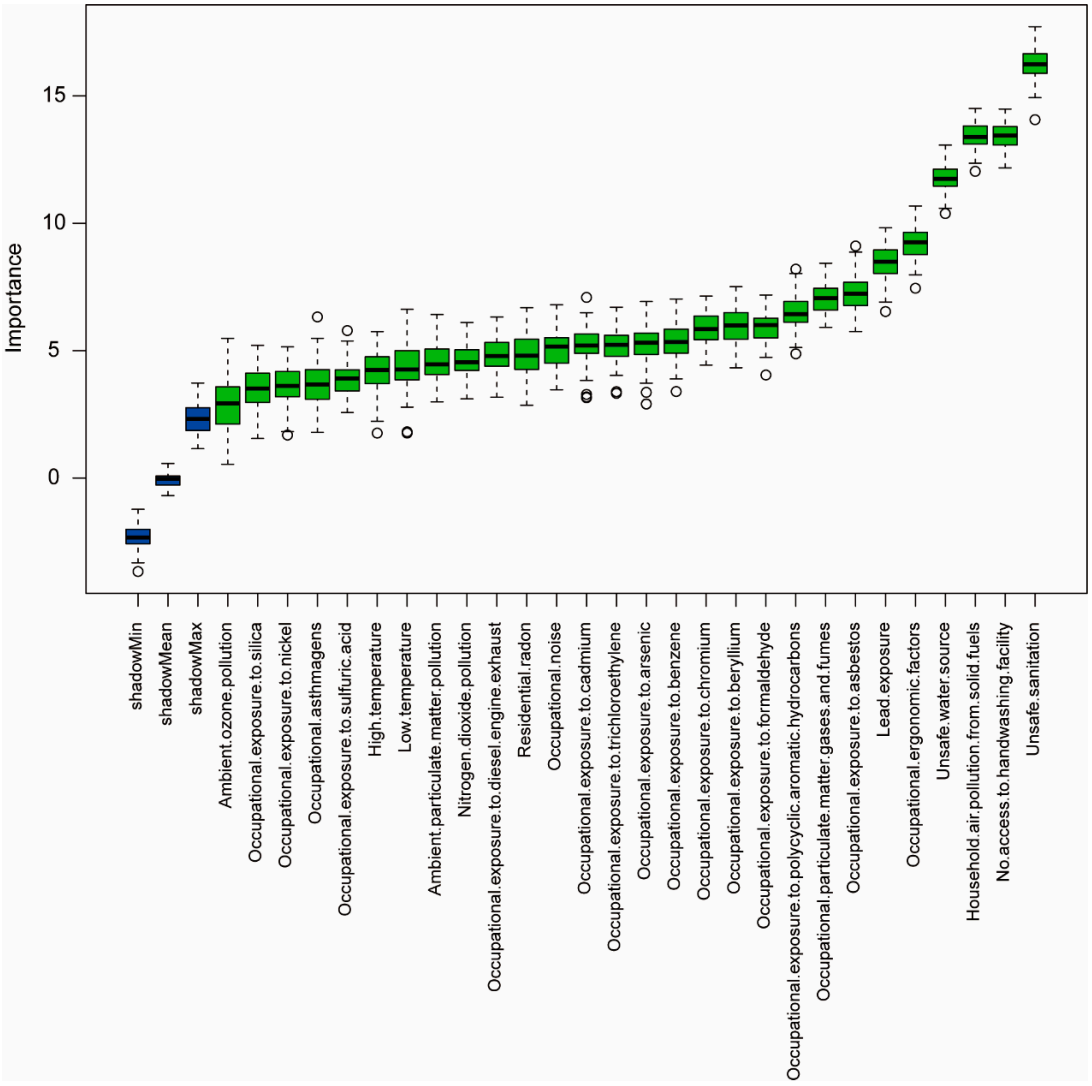

Figure 1: Importance scores of top candidate risk factors based on the Boruta algorithm.

Figure 2: Shapley additive explanations (SHAP) summary plot for the 28 Boruta-selected factors.

The SHAP value on the horizontal axis represents the influence of the factor on the outcome, with positive values indicating promotion and negative values indicating inhibition. The top 7 variables showed consistently positive contributions to congenital heart disease (CHD) risk.

Subsequently, we further quantitatively evaluated the association between the 7 major risk factors identified by the SHAP analysis and the disease risk. The 2021SDI was also included as an important potential confounder in the model to account for the influence of differences in socioeconomic development levels across countries or regions. First, we fitted a Poisson regression model using the same dependent variable, offset term, and covariates. Overdispersion was assessed using the Pearson dispersion parameter. The results showed that the Pearson dispersion parameter was 546.485, and the corresponding test was highly significant ($p < 0.001$), indicating substantial overdispersion. These findings demonstrated that the

assumptions of the Poisson model were violated and that the negative binomial regression model was therefore more appropriate to calculate the incidence rate ratio (IRR) and 95% confidence interval (95% CI) for the present data. Among the 7 risk factors and confounders, binomial regression showed that unsafe sanitation was the most significant independent risk factor, with an IRR of 1.48 (95% CI: 1.30–1.68), outperforming other environmental or occupational exposure factors (Fig. 3). This finding suggested that the lack of sanitation facilities may be associated with higher disease burden through multiple pathways such as pathogen exposure, intestinal infection, and immune stress, and is an important controllable risk source. After adjusting for SDI, other risk factors such as household air pollution from solid fuels, no access to handwashing facility, unsafe water source, lead exposure, occupational particulate matter, gases and fumes, and high temperature had little effect on the occurrence of CHD with IRR values close to 1 and 95% CIs crossing 1, suggesting that their risk contributions were relatively small and the effects were uncertain. The comprehensive analysis results indicated that unsafe sanitation is the most significant independent risk factor in this study population, and its impact intensity is much higher than that of other environmental or occupational exposure factors.

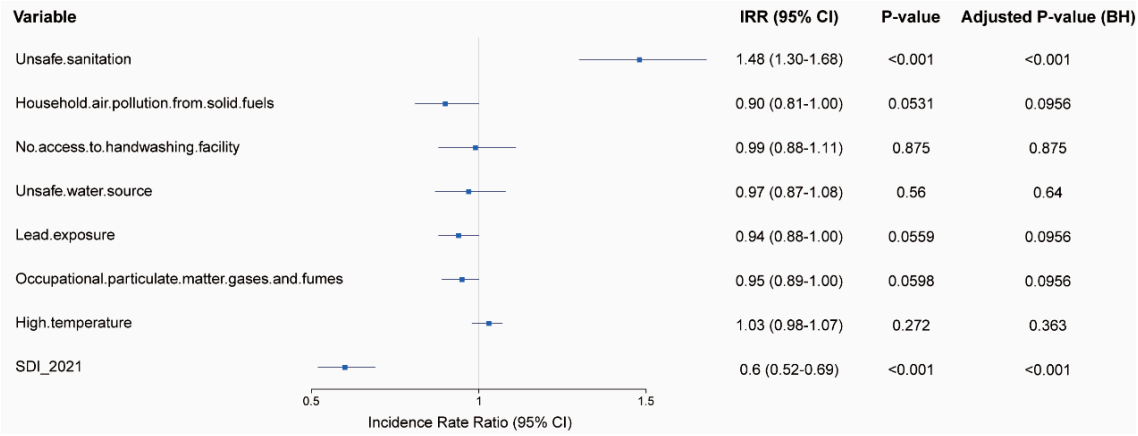

Figure 3: Forest plot and incidence rate ratio (IRR) values of major risk factors and confounders.

Negative binomial regression was used to estimate incidence rate ratio (IRR) values for the seven major risk factors identified by SHAP, adjusting for the 2021 Sociodemographic Index (SDI). Values of 7 major risk factors identified by SHAP and the 2021 sociodemographic index (SDI) were obtained by negative binomial regression independently conducted in this study. Adjusted *p*-values were calculated using the Benjamini-Hochberg (BH) procedure.

3.2 Global Spatial Distribution and Clustering Patterns of CHD and Unsafe Sanitation

Next, to reveal the spatial distribution characteristics of CHD incidence and its main risk factor, unsafe sanitation, this study first produced the global quartile distribution map for both. The spatial distribution demonstrated marked geographic heterogeneity, with substantial overlap between regions exhibiting high CHD incidence and high unsafe sanitation exposure, suggesting spatial concordance between these variables (Fig. 4).

To further explore the spatial distribution characteristics of CHD and its spatial correlation with environmental risk factors, this study conducted a global spatial autocorrelation analysis of CHD incidence rates and key risk factors. The results showed that the Moran's I value for CHD was 0.68342 (p = 0.001), indicating a significant spatial positive autocorrelation distribution feature within the study area. That is,

areas with higher incidence rates tend to be spatially adjacent, forming obvious high-value aggregation areas (hot spots), while areas with lower incidence rates also show a tendency to be adjacent to each other, indicating that the disease distribution is not random but has a certain spatial clustering property. At the same time, Moran's I value of unsafe sanitation was 0.72347 (p = 0.001), also showing a significant spatial clustering feature. Its spatial distribution pattern was highly similar to that of CHD, suggesting that the uneven distribution of environmental sanitation conditions may have a geographical overlap or spatial coupling relationship with the high-incidence areas of the disease. This result indicated potential spatial correlation between unsafe sanitation and the incidence rate of CHD, and the differences in regional health infrastructure construction may affect disease risk through environmental exposure pathways.

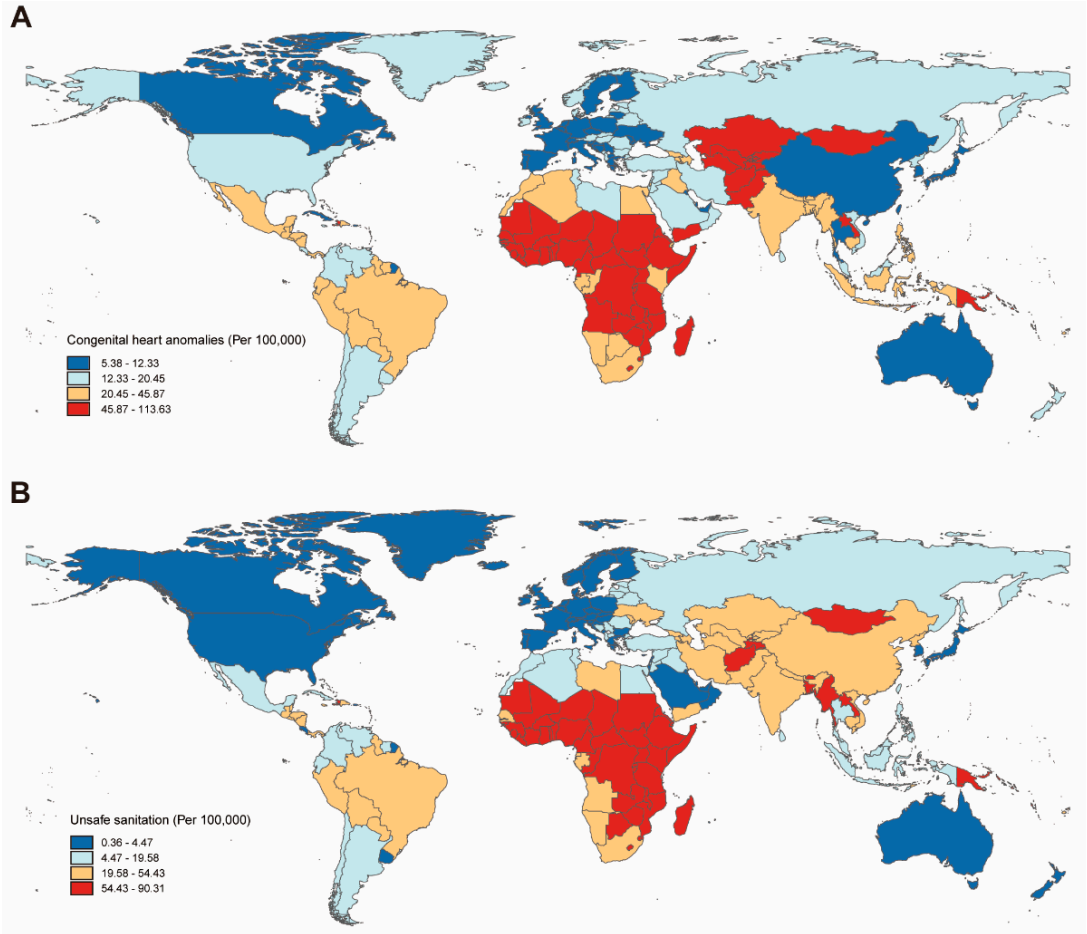

Figure 4: Global distribution of CHD incidence and unsafe sanitation exposure. (**A**) Quartile distribution of the incidence of congenital heart anomalies across 204 countries and territories. (**B**) Quartile distribution of the SEV of unsafe sanitation across 204 countries and territories.

In the subsequent LISA local spatial autocorrelation analysis, both exhibited significant spatial clustering. Spatial autocorrelation analyses consistently demonstrated that both CHD incidence and unsafe sanitation exhibited significant clustering, supporting the existence of shared geographic patterns rather than random spatial distributions (Fig. 5). This consistency across geographic space suggested that environmental hygiene conditions may play an important role in regional differences in incidence. The weak health infrastructure, low levels of clean water and sewage treatment in African and South Asian countries may

significantly amplify the risk of CHD through infection during pregnancy, malnutrition, and environmental exposure pathways.

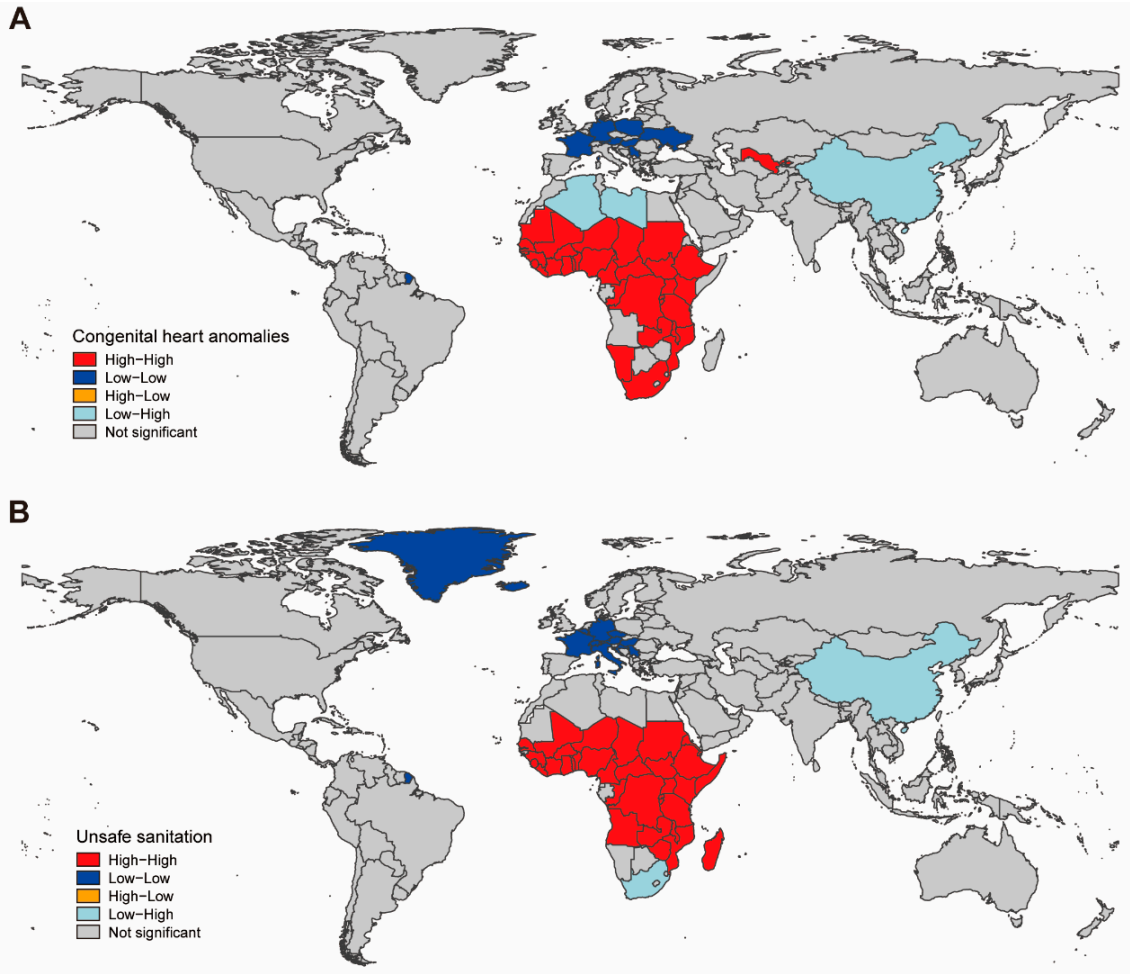

Figure 5: Spatial autocorrelation of CHD incidence and unsafe sanitation. (**A**) Local Indicators of Spatial Association (LISA) cluster maps of the incidence of congenital heart anomalies across 204 countries and territories. (**B**) LISA cluster maps of the SEV of unsafe sanitation across 204 countries and territories.

3.3 Attributable Burden and Socioeconomic Gradient of Unsafe Sanitation

On a global scale, the PAF of unsafe sanitation for CHD showed significant spatial heterogeneity. Approximately 33% of the CHD burden could theoretically be attributable to unsafe sanitation under the assumptions of the present statistical model. From a national perspective, the PAF was generally lower in high-income countries, mostly below 10%, such as North America, Western Europe, Australia, and Japan (Fig. 6A). In contrast, the PAF in sub-Saharan Africa, South Asia, and some Southeast Asian countries was generally above 40%, with many countries exceeding 60%. This reflected severe deficiencies in health infrastructure, such as in Chad, Niger, South Sudan, Somalia, and Malawi, indicating that children in these areas were exposed to a very high risk of CHD due to unsafe sanitation. More importantly, there was a significant negative correlation between national PAF and the SDI (Fig. 6B). The linear fitting results suggested that nearly 80% of the differences in PAF can be explained by the national socio-economic development level. This strong correlation indicated that the construction of sanitation facilities was highly constrained by

economic development, and unsafe sanitation may be an important mediating factor connecting poverty and CHD risk. As SDI increases, countries have achieved significant improvements in water supply, drainage, sewage treatment, toilet coverage, and hygiene education, resulting in a significant reduction in risks such as pathogen exposure, parasitic infection, water pollution, and exposure to environmental toxins, thereby simultaneously reducing the environmental burden of CHD. In high PAF areas, there were also "High-High" clusters of CHD, indicating a significant spatial overlap between exposure and outcome.

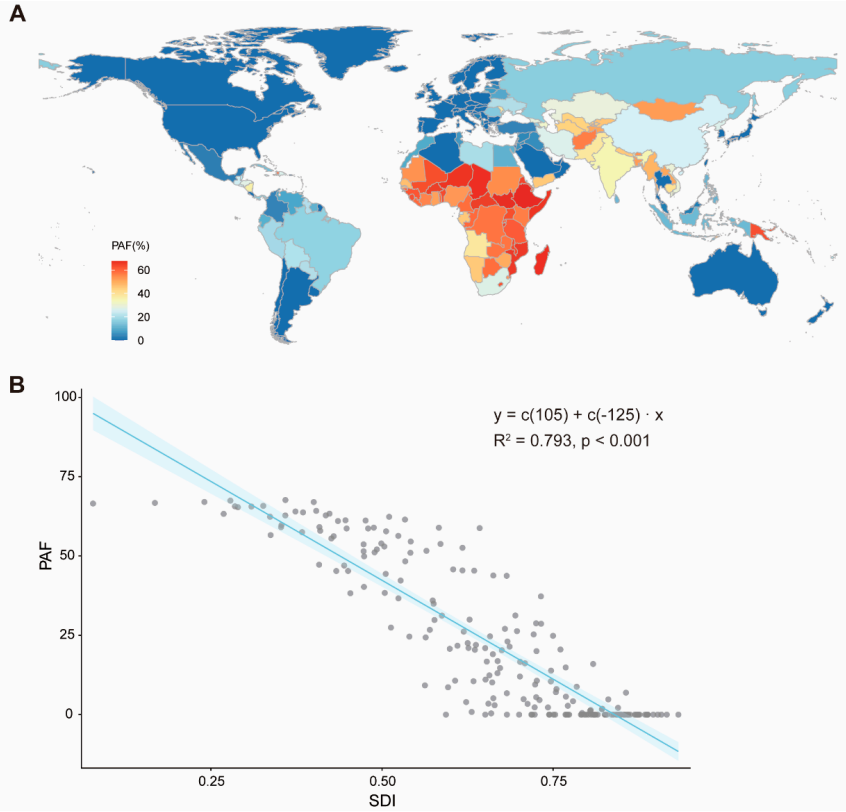

Figure 6: PAF of unsafe sanitation and its relationship with SDI. (**A**) Global distribution of population attributable fraction (PAF) for CHD attributable to unsafe sanitation. (**B**) Regression analysis of PAF and social demographic indices (SDI).

Overall, the distribution of global and regional PAFs was highly consistent with the previous spatial aggregation analysis results, further strengthening the potential relationship between unsafe sanitation conditions and the incidence of CHD. The evidence of spatial overlap between high PAF regions and high incidence regions suggested that improving sanitation infrastructure could potentially contribute to reducing the global burden of CHD. However, this hypothesis requires confirmation in prospective epidemiological and mechanistic studies.

4 Discussion

Based on the GBD 2021 database, we systematically analyzed global incidence patterns of CHD and the multidimensional environmental and behavioral risk factors that may influence it. Firstly, the Boruta algorithm was used to identify 28 potentially important factors, and then SHAP values were used to explain the directionality of the screening results, indicating that seven risk factors, such as unsafe sanitation, may be positively associated with CHD. Subsequently, this study used negative binomial regression to

quantitatively analyze seven significant factors and found that the incidence rate ratio of unsafe sanitation (IRR = 1.48, 95% CI: 1.30–1.68) was the highest, suggesting that unsafe sanitation is strongly associated with the global distribution of CHD and may represent an important environmental correlate deserving further investigation. Additionally, this study used Moran's I and LISA spatial autocorrelation analyses to find that CHD incidence exhibited significant spatial clustering worldwide, which overlapped with the spatial distribution pattern of unsafe sanitation. This spatial consistency provided geographical evidence supporting a potential link between this exposure factor and the disease. Further, through PAF analysis, this study quantified the contribution of unsafe sanitation to the burden of CHD, showing that the global PAF in 2021 was over 33%, and in several countries in sub-Saharan Africa and South Asia, the PAF was significantly over 60%, suggesting that unsafe sanitation exposure may be associated with a significant proportion of CHD cases in these regions. This study revealed the potential importance of exposure to unsafe sanitation in the occurrence of CHD through a data-driven, comprehensive evidence chain, providing new research perspectives and a scientific basis for future global prevention and control strategies for birth defects.

This study found that the incidence of CHD and unsafe sanitation conditions globally exhibits a highly similar spatial pattern. High incidence and high exposure are concentrated in sub-Saharan Africa and South Asia, while middle-income and developed regions show low values. This conclusion is highly consistent with previous large-scale epidemiological and public health monitoring reports. Multiple global and regional systematic analyses have pointed out that the burden of CHD has been consistently higher in low-middle income countries than in high-income countries, and the incidence or prevalence rates in Africa and South Asia are at a relatively high level globally [20,21]. For example, systematic analysis of global CHD and meta-analysis of birth prevalence rates both reported higher incidence burdens in developing countries [2,3]. At the same time, long-term monitoring by the World Health Organization and international institutions has shown that unsafe sanitation conditions are widespread in sub-Saharan Africa and South Asia, and are strongly associated with adverse health outcomes such as intestinal infections and malnutrition [13,22,23]. This is consistent with the high-exposure areas of unsafe sanitation in this study. However, there have been reports that some high or middle-income countries show relatively higher CHD reporting rates in individual studies, largely influenced by differences in diagnostic capabilities, completeness of birth registration, and screening policies for birth defects [3,19,24]. On the other hand, this study conducted ecological comparisons at the national or regional level, which may mask the heterogeneity at the sub-national level within a country, such as urban-rural differences, concentrated exposure in poverty-stricken areas [25–28], and some previous cohort or case-control studies have revealed the complex mechanisms of different risk factors at more detailed geographical or population levels, which may lead to inconsistencies to some extent [29,30].

The spatial concentration of unsafe sanitation is closely related to factors such as a country's level of economic development, geographic location, and climate. Low-income and middle-low-income countries, due to limited financial resources and insufficient technical and human capital, have difficulty building and maintaining centralized drainage and sewage treatment systems, resulting in a widespread lack of safe sanitation facilities in rural and poor communities [31]. This creates a high-risk zone that coincides with poverty rates and public investment capabilities. WHO monitoring indicates that the coverage of safe sanitation in regions such as sub-Saharan Africa and South Asia is significantly below the global average [23], where overlapping areas of high PAF for unsafe sanitation and high incidence of CHD are identified in this study. Geographical location also directly affects the accessibility and cost of infrastructure. Inland countries have scarce water resources, making the cost of constructing sanitation facilities high [32]. The degradation of forest ecosystems, such as tropical rainforests, will indirectly harm basic hygiene

conditions and water supply security, leading to an increase in unsafe sanitation rate in remote forest areas [33]. Islands, due to their boundedness, smallness, and isolation, have unignorable geographical inequalities in access to drinking water and sanitation compared to the mainland [34]. These factors together have led to a high geographic concentration of unsafe sanitation. In terms of climate, areas with frequent monsoons and floods and shallow groundwater levels are more prone to sewage overflow and contaminated drinking water, thereby exacerbating health risks [35–37]. Unsafe sanitation not only directly leads to the incidence and mortality of CHD, but also indirectly restrains economic growth by reducing labor productivity, increasing medical expenses, and inhibiting educational participation [13,38–40], making it difficult for vulnerable regions to escape the cycle of the health-development deadlock. Therefore, this study emphasizes the importance of establishing safe sanitation facilities to reduce the incidence of CHD and other diseases, thereby breaking this vicious cycle. However, it should be noted that the GBD definition of unsafe sanitation represents a composite environmental exposure rather than a single sanitation technology. The present findings cannot distinguish the relative effectiveness of individual interventions such as sewerage systems, improved pit latrines, composting toilets, or handwashing facilities. Future studies integrating more detailed infrastructure indicators are warranted to identify the most effective public health strategies.

The study suggests that improving environmental sanitation may be a key intervention to reduce the global burden of CHD. It should be emphasized that this study does not directly investigate these biological mechanisms. Instead, the following pathways represent biologically plausible hypotheses supported by previous experimental and epidemiological studies that may explain the observed spatial associations. Firstly, maternal infection is one of the most direct and well-documented pathways. Unsafe sanitation may significantly increase the overall burden of infections in the digestive and respiratory tracts, including infections caused by fulminant intestinal pathogens, bacterial urinary tract infections, and several viral diseases. Systemic or local infections occurring in the early stages of pregnancy may lead to maternal immune activation, fever, and the release of inflammatory mediators, which then affect embryonic heart layer differentiation and heart tube formation through the placenta [41]. For example, the rubella virus has a teratogenic effect on the fetal heart [42], and modern research also indicates that exposure to various viral and bacterial agents is associated with an increased risk of CHD [43,44]. Secondly, chronic gastrointestinal infections and recurrent diarrhea can lead to maternal nutrient absorption disorders and deficiencies of micronutrients. Women exposed to unsafe sanitation for a long time are more likely to suffer from malnutrition and insufficient intake or absorption of essential nutrients such as vitamin B_{12}, zinc, and folic acid, while the one-carbon metabolism pathway plays a crucial role in embryonic heart development [45,46]. On the other hand, malnutrition weakens the maternal immune system, increases susceptibility to infection, and exacerbates inflammatory responses, further increasing the risk of CHD. Thirdly, an increasing number of recent studies have focused on the impact of the maternal microbiome on fetal development during pregnancy. Unsafe sanitation affects the metabolic and immune microenvironment at the maternal-fetal interface by altering the environmental microorganisms' exposure spectrum, promoting pathogens spread, and inducing maternal intestinal dysbacteriosis, thereby interfering with placental function, fetal nourishment, and organ formation. Several studies have shown that abnormal microbiota during pregnancy is associated with fetal developmental abnormalities, and the metabolic products of the maternal microbiota have potential effects on cardiovascular development during the critical period of embryonic development [47–49]. Finally, unsafe sanitation is often accompanied by chemical pollution of drinking water and the environment, including heavy metals, organic pollutants, and polycyclic aromatic hydrocarbons. These pollutants can come from sewage overflow and untreated discharges, or may accumulate in drinking water due to the small diameter of groundwater channels and the shallow aquifers'

susceptibility to contamination. Epidemiological studies and toxicological experiments have suggested that exposure to lead, arsenic, cadmium, mercury, and certain organic pollutants during pregnancy may increase the risk of congenital heart malformations through pathways such as increased oxidative stress, interference with embryonic cell signaling pathways, and induction of epigenetic changes [11,50,51]. In areas with poor sanitation, such chemical and microbial exposures often co-occur, potentially generating synergistic or additive effects, thereby further amplifying the negative impact on heart development. The spatial analysis results of this study indicated spatial clustering of unsafe sanitation, high PAF values, and a high degree of overlap with CHD incidence. These biological mechanisms provide multiple levels of rationality for causal explanations and also suggest that improving basic sanitation facilities has significant health economic benefits.

This study not only clarified the core role of unsafe sanitation in global spatial differences in CHD but also proposed a research framework that integrates machine learning and spatial epidemiology. This new approach provides new methodological paths and empirical evidence for global risk identification, mechanism exploration, and health resource investment and policy formulation for birth defects.

This study still has several limitations. Firstly, this study primarily relied on a cross-sectional ecological analysis framework, using countries or regions as statistical units. Although this macro-scale analysis could reveal global spatial and exposure patterns, it inevitably carries the risk of ecological fallacies and does not necessarily reflect the causal relationship between individual-level real exposure and outcomes. Secondly, this study relied on GBD-estimated data, including CHD incidence and exposure values for multiple risk factors. These data may be highly dependent on model calculations in some low-income countries due to data scarcity, thereby introducing uncertainty. In countries with limited healthcare infrastructure, incomplete birth registration, restricted access to prenatal echocardiography, or insufficient diagnostic capacity, CHD incidence may be underestimated. Such regional differences in data quality could introduce systematic bias into spatial analyses despite the standardized GBD estimation framework, and the more recent GBD 2023 data were not included into this analysis. Moreover, although machine-learning screening of risk factors helped identify potential associations, it still primarily reflected statistical correlations and may not reflect a true causal relationship. Thirdly, although this study combined various methods to form a more complete evidence chain, it still failed to account for the interactive effects of multiple factors such as genetic background, utilization rate of maternal health care, underlying maternal diseases, nutritional status, air pollution, and chemical pollution. Differences in these factors across countries may be an important component of the observed spatial differences. Moreover, sanitation infrastructure generally reflects relatively stable long-term environmental conditions rather than transient exposures. Consequently, the SEVs used in this study are assumed to approximate chronic environmental exposure among women of reproductive age. However, the ecological design cannot determine the precise timing of maternal exposure during the critical period of embryonic cardiac development. Unmeasured confounders and using the same dataset for variable selection and testing may lead to overfitting. Thus, these limitations suggest that future research needs to further combine individual-level pregnancy cohorts, biomarker measurements, refined exposure models, genomic datasets, and multi-level causal inference methods to more accurately identify the true causal relationship and biological pathways between unsafe sanitation and CHD.

5 Conclusions

Using an integrated framework combining machine learning and spatial epidemiology, this study identified unsafe sanitation as the environmental factor most strongly associated with the global incidence of CHD. Significant spatial concordance was observed between unsafe sanitation exposure and CHD burden

across countries, particularly in sub-Saharan Africa and South Asia. Nevertheless, these findings should be interpreted as ecological associations rather than causal evidence. The theoretical PAF estimate of approximately 33% is model-dependent and requires validation in individual-level prospective cohorts. Future prospective studies integrating individual-level environmental exposure, genetic susceptibility, and mechanistic investigations are warranted to validate these observations and inform targeted public health interventions.

Acknowledgement: The authors would like to thank all the subjects who participated in this study for their participation and cooperation. The authors would like to thank the GBD database.

Funding Statement: This evaluation study was supported by Peking University People's Hospital Scientific Research Development Funds (RDEB2024-13), Noncommunicable Chronic Diseases-National Science and Technology Major Project (2024ZD0539200), Peking University People's Hospital Talent Introduction and Research Startup Fund (2022-T-01), Beijing Science and Technology Rising Stars Cross-Project (20230484441), Science and Technology Project of Guizhou Province (KXJZ [2024]035), the Project of Guizhou Provincial Science and Technology Agency (KJLYRC-[2026]045), Shanghai Key Laboratory of Clinical Molecular Diagnostics for Pediatrics (20dz2260900), National Clinical Key Specialty Construction Project (10000015Z155080000004), the Project of National Natural Science Foundation of China (U25C2002).

Author Contributions: Zeye Liu and Jing Xie conceived and planned the study and supervised the analyses. Yi Shen, Zeye Liu, Zeyu Jing and Wenchuan Liao conducted the analyses and wrote the first draft. Yi Shi, Yiwei Liu, Xuanqi An, Yanjun Sun and Hao Zhang critically reviewed the integrity and plausibility of the data analysis. Yi Yan, Bei Feng, Yiwei Liu, Xuanqi An and Yi Shi revised the manuscript and were responsible for the integrity of data acquisition and statistical analyses. Yifan Zhu, Chenyu Jiang, Xingliang Zhou, Xu Huang, Tianyu Liu, Jian Liu and Yuxi Ji provided administrative, technical, and material support, supervision, and mentorship. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets can be obtained by contacting Hao Zhang (drzhanghao@126.com) and confirming that they will only be used for academic research purposes.

Ethics Approval: Not applicable.

Conflicts of Interest: We declare no competing interests. While, the funder of the study had no role in the study design, data collection, data analysis, data interpretation, or writing of the report.

References

1. Bai Z, Han J, An J, Wang H, Du X, Yang Z, et al. The global, regional, and national patterns of change in the burden of congenital birth defects, 1990–2021: an analysis of the global burden of disease study 2021 and forecast to 2040. EClinicalMedicine. 2024;77:102873. [CrossRef].
2. Zimmerman MS, Smith AGC, Sable CA, Echko MM, Wilner LB, Olsen HE, et al. Global, regional, and national burden of congenital heart disease, 1990–2017: a systematic analysis for the global burden of disease study 2017. Lancet Child Adolesc Health. 2020;4(3):185–200. [CrossRef].
3. van der Linde D, Konings EEM, Slager MA, Witsenburg M, Helbing WA, Takkenberg JJM, et al. Birth prevalence of congenital heart disease worldwide a systematic review and meta-analysis. J Am Coll Cardiol. 2011;58(21):2241–7. [CrossRef].
4. Xu J, Li Q, Deng L, Xiong J, Cheng Z, Ye C. Global, regional, and national epidemiology of congenital heart disease in children from 1990 to 2021. Front Cardiovasc Med. 2025;12:1522644. [CrossRef].
5. Deng Y, Tian R, Zhao J, Liu L, Sun P, Ping Z. Global survival trends in congenital heart disease: temporal, regional, and sex disparities from global burden of disease analyses. Int J Cardiol. 2025;443:133925. [CrossRef].
6. Choudhury TZ, Gilbert BL, Garg V. Genetic and environmental contributors to congenital heart disease. Curr Treat Options Cardiovasc Med. 2025;27(1):36. [CrossRef].

7. Boyd R, McMullen H, Beqaj H, Kalfa D. Environmental exposures and congenital heart disease. Pediatrics. 2022;149:e2021052151. [CrossRef].
8. Miller E, Cradock-Watson JE, Pollock TM. Consequences of confirmed maternal rubella at successive stages of pregnancy. Lancet. 1982;2(8302):781–4. [CrossRef].
9. Petersen I, Collings SL, McCrea RL, Nazareth I, Osborn DP, Cowen PJ, et al. Antiepileptic drugs prescribed in pregnancy and prevalence of major congenital malformations: comparative prevalence studies. Clin Epidemiol. 2017;9:95–103. [CrossRef].
10. Gomersall JC, Moore VM, Fernandez RC, Giles LC, Grzeskowiak LE, Davies MJ, et al. Maternal modifiable factors and risk of congenital heart defects: systematic review and causality assessment. BMJ Open. 2024;14(8):e082961. [CrossRef].
11. Jin S, Yoon SZ, Choi YJ, Kang G, Choi SU. Prenatal exposure to air pollutants and the risk of congenital heart disease: a Korean national health insurance database-based study. Sci Rep. 2024;14(1):16940. [CrossRef].
12. Wróblewski M, Miłek J, Godlewski A, Wróblewska J. The impact of arsenic, cadmium, lead, mercury, and thallium exposure on the cardiovascular system and oxidative mechanisms in children. Curr Issues Mol Biol. 2025;47(7):483. [CrossRef].
13. World Health Organization. Sanitation 2024 [Internet]. [cited 2026 Jun 21]. Available from: https://www.who.int/news-room/fact-sheets/detail/sanitation.
14. Patel R, Gupta A, Chauhan S, Bansod DW. Effects of sanitation practices on adverse pregnancy outcomes in India: a conducive finding from recent Indian demographic health survey. BMC Pregnancy Childbirth. 2019;19(1):378. [CrossRef].
15. Kmush BL, Walia B, Neupane A, Frances C, Mohamed IA, Iqbal M, et al. Community-level impacts of sanitation coverage on maternal and neonatal health: a retrospective cohort of survey data. BMJ Glob Health. 2021;6(10):e005674. [CrossRef].
16. Shi QY, Zhang JB, Mi YQ, Song Y, Ma J, Zhang YL. Congenital heart defects and maternal fever: systematic review and meta-analysis. J Perinatol. 2014;34(9):677–82. [CrossRef].
17. Cameron L, Chase C, Contreras Suarez D. Relationship between water and sanitation and maternal health: evidence from Indonesia. World Dev. 2021;147:105637. [CrossRef].
18. Stevens GA, Alkema L, Black RE, Boerma JT, Collins GS, Ezzati M, et al. Guidelines for accurate and transparent health estimates reporting: the GATHER statement. PLoS Med. 2016;13(6):e1002056. [CrossRef].
19. Brauer M, Roth G, Aravkin A, Zheng P, Abate K, Abate Y, et al. Global burden and strength of evidence for 88 risk factors in 204 countries and 811 subnational locations, 1990–2021: a systematic analysis for the global burden of disease study 2021. Lancet. 2024;403(10440):2162–203. [CrossRef].
20. Duan J, Ding R, Yu Y, Li M, Ruan Y, Hu Y, et al. Global and regional burden of congenital birth defects, 1990–2021: persistent healthcare disparities and emerging challenges from non-fatal health burden. BMJ Public Health. 2025;3(1):e001608. [CrossRef].
21. Zhang X, Feng Y, Ren J, Jin X, Li J, Hou Y, et al. Global, regional, and national burden of congenital heart disease, 1990–2021: a systematic analysis for the global burden of disease study 2021. Eur J Pediatr. 2025;184(4):253. [CrossRef].
22. World Health Organization. Water, sanitation and hygiene: burden of disease [Internet]. [cited 2026 Jun 27]. Available from: https://www.who.int/data/gho/data/themes/topics/water-sanitation-and-hygiene-burden-of-disease.
23. UNICEF Data. Progress on household drinking water, sanitation and hygiene 2000–2024: special focus on inequalities [Internet]. [cited 2026 Jun 27]. Available from: https://data.unicef.org/resources/jmp-report-2025/.
24. Liu Y, Chen S, Zühlke L, Black GC, Choy MK, Li N, et al. Global birth prevalence of congenital heart defects 1970–2017: updated systematic review and meta-analysis of 260 studies. Int J Epidemiol. 2019;48(2):455–63. [CrossRef].
25. Minhas AMK, Wyand RA, Ariss RW, Nazir S, Jain V, Al-Kindi SG, et al. Rural-urban trends in congenital heart disease-related mortality in the United States, 1999 to 2019. JACC Adv. 2022;1(2):100030. [CrossRef].
26. Miao Q, Dunn S, Wen SW, Lougheed J, Reszel J, Lavin Venegas C, et al. Neighbourhood maternal socioeconomic status indicators and risk of congenital heart disease. BMC Pregnancy Childbirth. 2021;21(1):72. [CrossRef].

27. Peyvandi S, Baer RJ, Chambers CD, Norton ME, Rajagopal S, Ryckman KK, et al. Environmental and socioeconomic factors influence the live-born incidence of congenital heart disease: a population-based study in California. J Am Heart Assoc. 2020;9(8):e015255. [CrossRef].
28. Kucik JE, Nembhard WN, Donohue P, Devine O, Wang Y, Minkovitz CS, et al. Community socioeconomic disadvantage and the survival of infants with congenital heart defects. Am J Public Health. 2014;104(11):e150–7. [CrossRef].
29. Dolk H, McCullough N, Callaghan S, Casey F, Craig B, Given J, et al. Risk factors for congenital heart disease: the Baby Hearts Study, a population-based case-control study. PLoS One. 2020;15(2):e0227908. [CrossRef].
30. Guo L, Zhao D, Zhang R, Li S, Liu R, Wang H, et al. A matched case-control study on the association between colds, depressive symptoms during pregnancy and congenital heart disease in northwestern China. Sci Rep. 2019;9(1):589. [CrossRef].
31. Dangui K, Jia S. Water infrastructure performance in sub-Saharan Africa: an investigation of the drivers and impact on economic growth. Water. 2022;14(21):3522. [CrossRef].
32. Adugna D. Challenges of sanitation in developing counties—evidenced from a study of fourteen towns, Ethiopia. Heliyon. 2023;9(1):e12932. [CrossRef].
33. Mapulanga AM, Naito H. Effect of deforestation on access to clean drinking water. Proc Natl Acad Sci U S A. 2019;116(17):8249–54. [CrossRef].
34. Della Tonia Marchesi M, do Couto EV, Gomes UAF. Geographical inequalities in access to water and sanitation among Brazilian maritime islands' inhabitants. Appl Geogr. 2023;161:103120. [CrossRef].
35. Lawrence Grant S, Lange S, Almeida S, Hoque B, Kjær Mackie Jensen P. Influence of seasonal hazards on water, sanitation, and hygiene-related behavior and implications for cholera transmission in Bangladesh. Am J Trop Med Hyg. 2023;108(3):518–23. [CrossRef].
36. Mohammed JN, Swalaha FM. Sewage intrusion into drinking water distribution systems: implications for water resource management. Environ Monit Assess. 2025;197(10):1150. [CrossRef].
37. Rahim FB, Swarnokar SC, Roy S, Rahman MM, Das TK. The quality of groundwater in shallow aquifers impacted by inadequately managed sanitation facilities in coastal Bangladesh. Clean Water. 2025;4:100154. [CrossRef].
38. Hutton G, Haller L, Bartram J. Global cost-benefit analysis of water supply and sanitation interventions. J Water Health. 2007;5(4):481–502. [CrossRef].
39. Sharma MK, Adhikari R, Khanal SP, Acharya D, van Teijlingen E. Do school water, sanitation, and hygiene facilities affect students' health status, attendance, and educational achievements? A qualitative study in Nepal. Health Sci Rep. 2024;7(8):e2293. [CrossRef].
40. Trinies V, Garn JV, Chang HH, Freeman MC. The impact of a school-based water, sanitation, and hygiene program on absenteeism, diarrhea, and respiratory infection: a matched-control trial in Mali. Am J Trop Med Hyg. 2016;94(6):1418–25. [CrossRef].
41. Wang S, Mei Z, Chen J, Zhao K, Kong R, McClements L, et al. Maternal immune activation: implications for congenital heart defects. Clin Rev Allergy Immunol. 2025;68(1):36. [CrossRef].
42. Mekonnen D. Clinically confirmed congenital rubella syndrome: the role of echocardiography. Ethiop J Health Sci. 2017;27(2):197–202. [CrossRef].
43. Su H, Guo E, Woodward M, He JR, Waterboer T, Schuermans A, et al. First trimester maternal infections and offspring congenital heart defects: a meta-analysis. Eur Heart J. 2026;47(7):794–812. [CrossRef].
44. Wang T, Li Q, Chen L, Ni B, Sheng X, Huang P, et al. Maternal viral infection in early pregnancy and risk of congenital heart disease in offspring: a prospective cohort study in Central China. Clin Epidemiol. 2022;14:71–82. [CrossRef].
45. Qu Y, Liu X, Lin S, Bloom MS, Wang X, Li X, et al. Maternal serum folate during pregnancy and congenital heart disease in offspring. JAMA Netw Open. 2024;7(10):e2438747. [CrossRef].
46. Qu Y, Lin S, Zhuang J, Bloom MS, Smith M, Nie Z, et al. First-trimester maternal folic acid supplementation reduced risks of severe and most congenital heart diseases in offspring: a large case-control study. J Am Heart Assoc. 2020;9(13):e015652. [CrossRef].
47. Zych-Krekora K, Sylwestrzak O, Krekora M, Oszukowski P, Grzesiak M. Potential impact of a pregnant woman's microbiota on the development of fetal heart defects: a review of the literature. Biomedicines. 2024;12(3):654. [CrossRef].

48. Barrientos G, Ronchi F, Conrad ML. Nutrition during pregnancy: influence on the gut microbiome and fetal development. Am J Reprod Immunol. 2024;91(1):e13802. [CrossRef].
49. Ruiz-Triviño J, Álvarez D, Cadavid JÁP, Alvarez AM. From gut to placenta: understanding how the maternal microbiome models life-long conditions. Front Endocrinol. 2023;14:1304727. [CrossRef].
50. Gorini F, Tonacci A. Toxic metals in pregnancy and congenital heart defects. Insights and new perspectives for a technology-driven reduction in food sources. Explor Cardiol. 2023;1(3):114–40. [CrossRef].
51. Liu Z, He C, Chen M, Yang S, Li J, Lin Y, et al. The effects of lead and aluminum exposure on congenital heart disease and the mechanism of oxidative stress. Reprod Toxicol. 2018;81:93–8. [CrossRef].