

目录

| | |
|---|----|
| Association of Short-Term Co-Exposure to Particulate Matter and Ozone with Mortality Risk | 1 |
| Blood multiple heavy metals exposure and lung function in young adults: A prospective Cohort study in China | 11 |
| Association of psychological distress and DNA methylation: A 5-year longitudinal population-based twin study | 19 |
| A high-resolution haplotype-resolved Reference panel constructed from the China Kadoorie Biobank Study | 28 |
| Immune-Boosting Effect of the COVID-19 Vaccine Real World Bidirectional Cohort Study | 41 |
| Improving_Cardiovascular_Risk_Prediction_through_Machine_Learning_Modelling_of_Irregular_Repeated_Electronic_Health_Records | 55 |
| Transmissibility quantification of norovirus outbreaks in 2016–2021 in Beijing, China | 85 |
| Spatiotemporal cluster of mpox in men who have sex with men: A modeling study in 83 countries | 97 |

Association of Short-Term Co-Exposure to Particulate Matter and Ozone with Mortality Risk

Published as part of the *Environmental Science & Technology virtual special issue* “Accelerating Environmental Research to Achieve Sustainable Development Goals”.

Jianhui Guo, Jinyi Zhou, Renqiang Han, Yaqi Wang, Xinyao Lian, Ziqi Tang, Jin Ye, Xueqiong He, Hao Yu,* Shaodan Huang,* and Jing Li*



Cite This: <https://doi.org/10.1021/acs.est.3c04056>



Read Online

ACCESS |



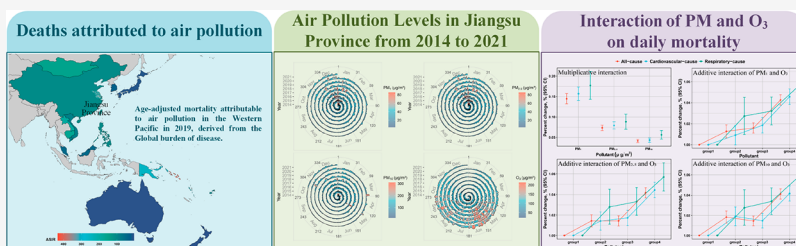
Metrics & More



Article Recommendations



Supporting Information



ABSTRACT: A complex regional air pollution problem dominated by particulate matter (PM) and ozone (O₃) needs drastic attention since the levels of O₃ and PM are not decreasing in many parts of the world. Limited evidence is currently available regarding the association between co-exposure to PM and O₃ and mortality. A multicounty time-series study was used to investigate the associations of short-term exposure to PM₁, PM_{2.5}, PM₁₀, and O₃ with daily mortality from different causes, which was based on data obtained from the Mortality Surveillance System managed by the Jiangsu Province Center for Disease Control and Prevention of China and analyzed via overdispersed generalized additive models with random-effects meta-analysis. We investigated the interactions of PM and O₃ on daily mortality and calculated the mortality fractions attributable to PM and O₃. Our results showed that PM₁ is more strongly associated with daily mortality than PM_{2.5}, PM₁₀, and O₃, and percent increases in daily all-cause nonaccidental, cardiovascular, and respiratory mortality were 1.37% (95% confidence interval (CI), 1.22–1.52%), 1.44% (95% CI, 1.25–1.63%), and 1.63% (95% CI, 1.25–2.01%), respectively, for a 10 μg/m³ increase in the 2 day average PM₁ concentration. We found multiplicative and additive interactions of short-term co-exposure to PM and O₃ on daily mortality. The risk of mortality was greatest among those with higher levels of exposure to both PM (especially PM₁) and O₃. Moreover, excess total and cardiovascular mortality due to PM₁ exposure is highest in populations with higher O₃ exposure levels. Our results highlight the importance of the collaborative governance of PM and O₃, providing a scientific foundation for pertinent standards and regulatory interventions.

KEYWORDS: Particulate matter, Ozone, Mortality, Interaction, Excess fraction

1. INTRODUCTION

Air pollution is widely recognized as a risk factor for human health.¹ In 2019, environmental pollution caused 9 million premature deaths, with air pollution being the primary cause.² Increasing conclusive evidence has demonstrated the positive associations of short-term exposure to air pollutants with all-cause and cause-specific mortality.^{3–5} As research continues, synergistic control of multiple pollutants (especially PM and O₃) is gaining widespread attention and is more effective than single pollutant control strategies.⁶ In many parts of the world, the phenomenon of PM–O₃ complex pollution has gradually come to the forefront and has become a major feature of air pollution.^{7–9} Therefore, attention to the synergistic control of PM and O₃ is essential to reduce the burden of disease.

However, few studies have examined the association between short-term co-exposure to PM and O₃ on daily mortality. Most studies only considered O₃ as a confounding factor when examining the associations of short-term exposure to PM with adverse health outcomes. Even so, some studies have found a stronger association between exposure to oxidizing gases and mortality risk in areas where the PM components have a greater capacity for oxidative stress (e.g.,

Received: May 29, 2023

Revised: September 15, 2023

Accepted: September 18, 2023



higher excess metal content and oxidation potential)^{10,11} and a synergistic effect of co-exposure to PM and O₃ through mechanisms such as neuro-immune interactions¹² and the production of inflammation.¹³ However, a meta-analysis on a number of previous studies found uncertainty about how short-term co-exposure to PM and O₃ exerts a combined effect to lead to adverse outcomes due to inconsistencies in the end points examined, the study design, the sample size, and the way exposure levels of pollutants are estimated (i.e., the composition of pollutants and the particle size of pollutants).¹⁴ For example, epidemiologic studies have shown that short-term co-exposure to PM and O₃ has a modifying effect on cardiovascular health rather than a combined effect,^{15,16} while experimental studies have also shown the presence of synergistic effects, antagonistic effects, or no overall interaction between PM and O₃.¹⁴ Therefore, there is an urgent need to further investigate whether there is a combined or interactive effect of short-term co-exposure to PM and O₃ on mortality using large population-based surveys.

The present study aims to examine the association of short-term co-exposure to PM and O₃ on all-cause nonaccidental, cardiovascular, and respiratory mortality in 4,276,989 people in Jiangsu Province, China, during 2014–2021. We hypothesized an interaction of short-term co-exposure to PM (especially PM₁) and O₃ on daily mortality. Our findings will provide a basis for the synergistic regulation of PM and O₃ and influence the development of relevant policies.

2. METHODS

2.1. Study Area and Population. Jiangsu Province is located on the east coast of China with a total area of 107,200 km². There are 13 prefecture-level cities and 96 counties/districts (termed “counties”), and the province had a population of 85.1 million by the end of 2021. Jiangsu Province is also one of the most developed provinces in China. In recent years, the health burden of severe air pollution in Jiangsu Province has caused widespread public concern due to rapid industrial development and urbanization. Our study population was from the Mortality Surveillance System managed by the Jiangsu Province Center for Disease Control and Prevention of China. We collected information about the cause of death, date of death, address of usual residence, and socio-demographic characteristics for each of the 4,276,989 participants who died from nonaccidental causes in Jiangsu Province, China, during 2014–2021, including 1,631,135 participants who died from cardiovascular system diseases and 489,730 participants who died from respiratory system diseases.

2.2. Outcomes. We calculated county-level, daily, all-cause nonaccidental, cardiovascular, and respiratory mortality in Jiangsu Province during 2014–2021. The underlying cause of death was classified according to the 10th revision of the International Classification of Diseases (ICD) code. All-cause nonaccidental disease was categorized as A00–R99, including intentional self-harm (X60–X84), which is known to be influenced by short-term air pollution.¹⁷ Cardiovascular disease was classified as I00–I99, and respiratory disease was classified as J00–J98.

2.3. Exposure Assessment. We retrieved the daily high spatial-temporal PM₁,¹⁸ PM_{2.5},¹⁹ and PM₁₀²⁰ exposure data for Jiangsu Province in 2013–2021 from the China High Air Pollutant (CHAP) data set (spatial resolution 1 km × 1 km), available at <https://weijing-rs.github.io/product.html>. We also

retrieved the daily gridded O₃ data for Jiangsu Province in 2013–2021 from the Tracking Air Pollution (TAP) in China data set (spatial resolution 10 km × 10 km).²¹ These data are in good agreement with the results from the ground-based detection stations. The coefficients of determination for cross-validation were 0.83 (PM₁), 0.92 (PM_{2.5}), 0.90 (PM₁₀), and 0.70 (O₃). We extracted the 24 h average PM₁, PM_{2.5}, and PM₁₀ and the maximum 8 h moving average O₃ concentrations for 96 counties in Jiangsu Province. We also extracted temperature and humidity data for each county in Jiangsu Province from the ERAS-Land data set.²² Figure S1 and Table S1 show the average annual and daily levels of air pollutants, temperature, and humidity for Jiangsu Province in 2014–2021.

2.4. Statistical Analysis. **2.4.1. Single Effects.** We used a two-stage analytic protocol widely used for multicounty time-series studies using the R packages “mgcv” and “metafor” to explore the associations between short-term exposure to air pollution and mortality.^{3,23} In the first stage, county-specific linear associations of PM (including PM₁, PM_{2.5}, and PM₁₀) and O₃ exposures with mortality were estimated using a generalized additive model with quasi-Poisson regression.^{3,23} Due to the small number of mortality from respiratory diseases in each county, associations were estimated in the first stage using zero-inflated Poisson regression in order to avoid overfitting situations.²⁴ In our model, we tried PM and O₃ exposures at different windows. We determined exposure windows using a generalized cross-validation (GCV) score, which was used for subset selection of regression and singular value truncation methods as well as for best model selection.²⁵ We also included these factors in our model: (1) the long-term trend, or natural cubic spline functions with 7 degrees of freedom (df) per year; (2) an indicator variable for “day of week” to account for possible variations in a week; (3) 4 day (lag 0–3 days) average temperatures, or natural cubic spline functions with 6 df;^{26–28} (4) 4 day (lag 0–3 days) average humidity, or natural cubic spline functions with 3 df for 4 days of humidity;^{26–28} and (5) the season, including the cold season (January–March and December) and the warm season (April–November).

In the second stage, a random-effects meta-analysis was used to pool the estimates of the county-specific associations. This analysis has been widely used in multiregional epidemiological studies, taking into account both intraregional statistical errors and interregional variability. We reported the pooled estimates with a 95% confidence interval (CI), which is the percent change in mortality corresponding to every 10 μg/m³ increase in PM and O₃ concentrations.

Sensitivity analyses were performed to validate the robustness of our results. We excluded the top 5% and bottom 5% of pollutant measurement data for each county and compared them to the results of our original model.²⁷ We also considered the 96 counties as a covariate in the generalized additive model and compared the results with a two-step analysis. Additionally, we performed stratified analyses by sex (male vs female), age (≤75 years vs >75 years), and season (warm season vs cold season) to search for susceptibility factors influencing the associations of short-term exposure to PM and O₃ with daily mortality. In all stratified analyses, we used a generalized additive model with zero-inflated Poisson regression in the first stage.

Moreover, we also obtained nonlinear exposure–response relationships for PM₁, PM_{2.5}, PM₁₀, and O₃ with mortality using the R packages “d lnm”, “splines”, and “mvmeta” by

replacing the linear terms for the pollutants with a natural spline function in the first-stage model. The natural spline function had three nodes at the 25th, 50th, and 75th percentiles of pollutant exposure.

2.4.2. Combined Effects and Interactions of Short-Term Exposure to PM and O₃ on Daily Mortality. We grouped the study participants based on the Air Quality Guidelines of the World Health Organization (WHO-AQG) for O₃ (maximum 8 h moving average concentrations), including a low O₃ exposure group (<100 μg/m³) and a high O₃ exposure group (≥100 μg/m³), to examine the impact of O₃ on the associations between short-term PM exposure and daily mortality. We also used the two-stage analysis described above to aggregate county-specific estimates of the multiplicative interaction of PM and O₃ on daily mortality.

In addition, we explored the additive interaction of short-term exposure to PM and O₃ on daily mortality. Specifically, we grouped the participants according to the WHO-AQG for PM₁₀ (24 h average concentrations, 45 μg/m³) and O₃ (100 μg/m³), the WHO Air Quality Interim Target 4 for PM_{2.5} (24 h average concentrations, 25 μg/m³) to avoid under-representation in one group, and the WHO-AQG of PM_{2.5} (15 μg/m³) for PM₁ due to the lack of current standards for PM₁. We examined the additive interaction between PM and O₃ on daily mortality by including each county as a covariate in a generalized linear model with a quasi-Poisson regression.

2.4.3. Evaluation of Excess Mortality. We calculated the excess mortality associated with short-term exposure to PM and O₃ in each county using a previously described two-stage analytic protocol and the R packages “d lnm”, “splines”, and “mixmeta”. Briefly, we calculated the excess daily deaths by cumulative RR within lag 0–1 using the standard formula $(1 - \exp(-\beta_j(x_{jt} - 0)_+)) \times d_{jt}$ for continuous exposure.^{29,30} In this formula, β_j represents the log RR for an increase of 10 μg/m³ in PM and O₃, respectively, defined as the county-specific best linear unbiased prediction for county j , and d_{jt} and x_{jt} are the corresponding daily mortality at day t and the average PM and O₃ levels in the same day and the day prior. The term $(x_{jt} - c)_+$ represents the excess PM and O₃ concentrations of the above the limit c . According to previous studies,^{3,31} 0 μg/m³ was chosen as the limit value for PM because there is no evidence of a threshold for the exposure–response relationship between PM and mortality, and 100 μg/m³ was chosen as the limit value for O₃. We reported fractions of excess deaths with 95% empirical confidence intervals (eCIs). We also grouped the study participants according to the WHO-AQG for O₃ to calculate the excess mortality associated with short-term PM₁, PM_{2.5}, or PM₁₀ exposure in each county.

All analyses were performed with R software (ver. 4.2.3; R Development Core Team). A p -value of less than 0.05 was considered to indicate statistical significance.

3. RESULTS

3.1. Study Population. In the study, a total of 4,276,989 participants were included, of whom 2,360,573 (55.2%) were men, 2,568,429 (60.1%) were aged 75 years or older, and 2,646,558 (61.9%) died during the warm season. A total of 1,631,135 study participants died of cardiovascular disease, of whom 811,287 (49.7%) were men, 1,177,471 (72.2%) were aged 75 years or older, and 973,653 (59.7%) died during the warm season. A total of 489,730 study participants died from respiratory diseases in this study, of whom 278,998 (57.0%) were men, 399,320 (81.5%) were aged 75 years or older, and

272,772 (55.7%) died during the warm season (Table 1). The total number of deaths and the average daily number of deaths

Table 1. Descriptive Characteristics of the Participants in This Study^a

| variables | all-cause mortality ($n = 4,276,989$) (%) | cardiovascular mortality ($n = 1,631,135$) (%) | respiratory mortality ($n = 489,730$) (%) |
|------------------------------|---|--|---|
| Sex | | | |
| male | 2,360,573 (55.2) | 811,287 (49.7) | 278,998 (57.0) |
| female | 1,916,416 (44.8) | 819,848 (50.3) | 210,732 (43.0) |
| Age (Year) | | | |
| 0–75 | 1,708,560 (39.9) | 453,664 (27.8) | 90,410 (18.5) |
| >75 | 2,568,429 (60.1) | 1,177,471 (72.2) | 399,320 (81.5) |
| Marital | | | |
| unmarried | 156,732 (3.7) | 40,291 (2.5) | 14,273 (2.9) |
| married | 2,712,315 (63.4) | 953,038 (58.4) | 268,226 (54.8) |
| divorced | 44,105 (1.0) | 14,475 (0.9) | 2974 (0.6) |
| widowed | 1,348,032 (31.5) | 617,565 (37.9) | 202,867 (41.4) |
| unspecified | 15,805 (0.4) | 5766 (0.3) | 1390 (0.3) |
| Education | | | |
| junior high school and below | 3,963,749 (92.7) | 1,535,982 (94.2) | 465,720 (95.1) |
| high school and above | 313,240 (7.3) | 95,153 (5.7) | 24,010 (4.8) |
| Season | | | |
| warm | 2,646,558 (61.9) | 973,653 (59.7) | 272,772 (55.7) |
| cold | 1,630,431 (38.1) | 657,482 (40.3) | 216,958 (44.3) |

^aData are presented as the absolute number (percentages) for the categorical variables.

for each county are shown in Table S2. The change in the number of daily deaths over time for the province in 2014–2021 is shown in Figure S2.

3.2. Associations of PM and O₃ with Daily Mortality.

The associations between air pollution exposure levels and daily all-cause mortality for different lag days of PM and O₃, respectively, showed that a 2 day average (the average exposure on the mortality day and the day prior (lag 0–1 days)) had the smallest mean generalized cross-validation score compared to other exposure windows. The association of PM and O₃ with daily all-cause mortality was stronger at this time (lag 0–1 days) (Table S3). The 4 day average temperature (including the temperature on the mortality day and the 3 days prior (lag 0–3 days)) was found to have produced the lower mean generalized cross-validation scores and was associated more strongly with daily all-cause deaths (Table S4). Based on the above analysis and literature findings,^{3,4} a 2 day average exposure to pollutants and a 4 day average temperature were used for subsequent analyses.

Finally, we found significant positive associations of PM₁, PM_{2.5}, PM₁₀, and O₃ with daily all-cause, cardiovascular, and respiratory mortality (Table 2). At the provincial level, we observed that a per 10 μg/m³ increase in PM₁ was associated with 1.37% (95% CI, 1.22–1.52%) in a pooled estimate of all-cause mortality, 1.44% (95% CI, 1.25–1.63%) in a pooled estimate of cardiovascular mortality, and 1.63% (95% CI, 1.25–2.01%) in a pooled estimate of respiratory mortality. We also observed that the associations of PM₁, PM_{2.5}, and PM₁₀ with daily mortality decreased with increasing particle size. Furthermore, the results suggest that a per 10 μg/m³ increase in O₃ was associated with 0.80% (95% CI, 0.73–0.87%) in a

Table 2. Percentage Change in All-Cause Mortality, Cardiovascular Mortality, and Respiratory Mortality per 10 $\mu\text{g}/\text{m}^3$ Increase in the 2 Day Moving Average Concentrations of PM_{10} , $\text{PM}_{2.5}$, PM_{10} , and O_3 at the Provincial Level

| pollutants | pooled estimate, % (95% CI) ^a | | |
|-------------------|--|--------------------------|-----------------------|
| | all-cause mortality | cardiovascular mortality | respiratory mortality |
| PM_{10} | 1.37 (1.22, 1.52) | 1.44 (1.25, 1.63) | 1.63 (1.25, 2.01) |
| $\text{PM}_{2.5}$ | 0.60 (0.52, 0.67) | 0.63 (0.53, 0.72) | 0.70 (0.51, 0.89) |
| PM_{10} | 0.36 (0.31, 0.41) | 0.37 (0.30, 0.43) | 0.50 (0.38, 0.62) |
| O_3 | 0.80 (0.73, 0.87) | 0.85 (0.75, 0.96) | 1.15 (0.94, 1.36) |

^aPooled estimates represent the percentage changes in daily all-cause, cardiovascular, and respiratory mortality per 10 $\mu\text{g}/\text{m}^3$ increase in the concentrations of particulate matter (PM) with an aerodynamic diameter of 1 μm or less (PM_{10}), 2.5 μm or less ($\text{PM}_{2.5}$), and 10 μm or less (PM_{10}) and ozone (O_3).

pooled estimate of all-cause mortality, 0.85% (95% CI, 0.75–0.96%) in a pooled estimate of cardiovascular mortality, and

1.15% (95% CI, 0.94–1.36%) in a pooled estimate of respiratory mortality.

The exposure–response relationships of PM and O_3 with daily mortality demonstrated a gradual increase in mortality with an increasing level of exposure to the pollutants. For PM, the slope of the curve is steeper at lower levels. Even in populations exposed to lower levels of PM pollution than the WHO standard, there were still detectable positive associations with mortality. For O_3 , the daily mortality showed a significant increase only at concentrations above 50 $\mu\text{g}/\text{m}^3$ (Figures 1, 2, and 3).

Stratified analysis showed that there were association modifiers for sex, age, and season in the association of PM and O_3 with daily mortality. Participants aged over 75 years and women were more susceptible to air pollution. The association of air pollution is stronger in the warm season (April–November) (Tables S7–S9).

We also conducted a sensitivity analysis by removing the top 5% and bottom 5% of pollutant exposure concentrations for each county. The results are consistent with those described previously, suggesting a significant association of PM_{10} , $\text{PM}_{2.5}$, PM_{10} , and O_3 with daily mortality (Table S5). In another

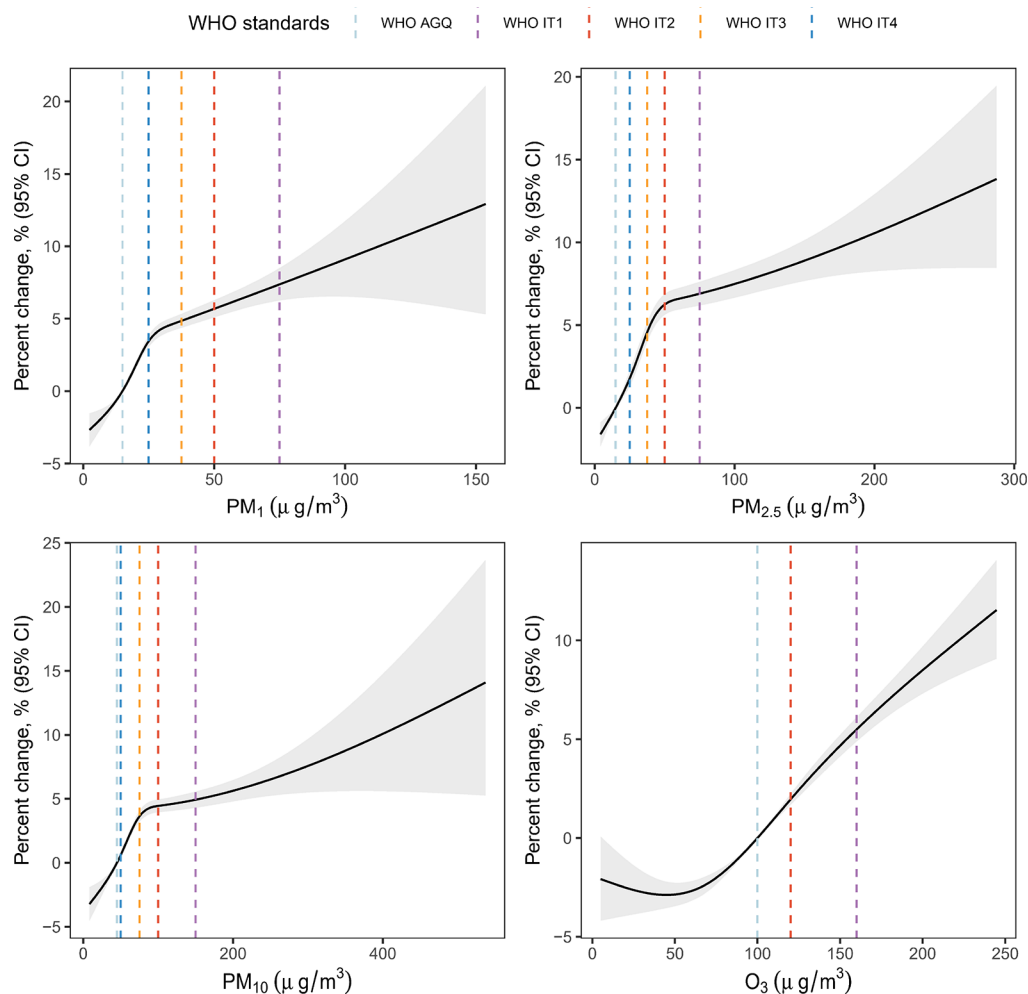


Figure 1. Pooled concentration–response curves for all-cause mortality. Shown are the pooled concentration–response curves for the associations of the 2 day moving average concentrations of PM_{10} , $\text{PM}_{2.5}$, PM_{10} , and O_3 with daily all-cause mortality. The y-axis displays the percentage difference in mortality from the pooled mean effect. The dashed lines indicate the air quality guidelines or standards for particulate matter (PM_{10} , $\text{PM}_{2.5}$, or PM_{10}) or ozone (8 h average concentrations) according to the World Health Organization (WHO) Air Quality Guidelines as well as four WHO Interim Targets (IT-1, IT-2, IT-3, and IT-4).

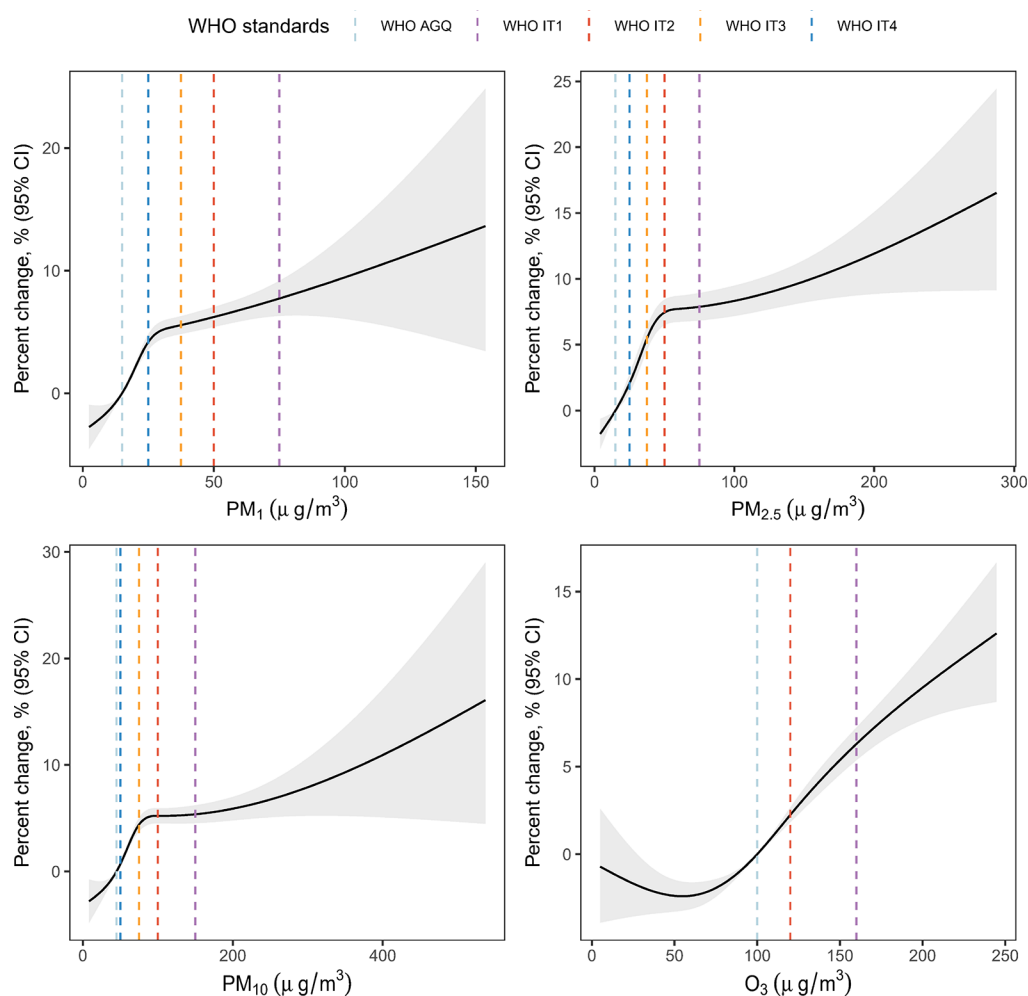


Figure 2. Pooled concentration–response curves for cardiovascular mortality. Shown are the pooled concentration–response curves for the associations of the 2 day moving average concentrations of PM_{10} , $PM_{2.5}$, PM_{10} , and O_3 with daily cardiovascular mortality. The y-axis displays the percentage difference in mortality from the pooled mean effect. The dashed lines indicate the air quality guidelines or standards for particulate matter (PM_{10} , $PM_{2.5}$, or PM_{10}) or ozone (8 h average concentrations) according to the World Health Organization (WHO) Air Quality Guidelines as well as four WHO Interim Targets (IT-1, IT-2, IT-3, and IT-4).

sensitivity analysis, which incorporated both the daily mortality and pollutant levels for each county into a generalized additive model with quasi-Poisson regression and treated county names as covariates, similar associations were found (Table S6).

3.3. Combined and Interactive Effects of PM and O_3 on Daily Mortality. We grouped the study participants according to O_3 exposure levels, and our results showed that those with high O_3 exposure levels had greater susceptibility to PM_{10} , $PM_{2.5}$, and PM_{10} . There was a significant multiplicative interaction between short-term exposure to PM (PM_{10} , $PM_{2.5}$, and PM_{10}) and O_3 on daily all-cause, cardiovascular, and respiratory mortality (Table 3). We also found a significant additive interaction of PM (PM_{10} and $PM_{2.5}$) and O_3 on daily all-cause, cardiovascular, and respiratory mortality. The combined effect showed that those with high exposure to both PM and O_3 had a significantly increased risk of death compared to those with low exposure to both (Figures S3–S5).

3.4. Excess Mortality Attributable to PM and O_3 . We calculated the excess mortality fraction (%) attributable to short-term exposure to PM and O_3 . The results showed that PM_{10} exposure was associated with an increased excess mortality fraction of all-cause, cardiovascular, and respiratory

disease by 3.76% (95% eCI, 3.65–3.86%), 4.06% (95% eCI, 3.93–4.18%), and 4.39% (95% eCI, 4.01–4.72%), respectively. Excess mortality from PM_{10} was higher than that from $PM_{2.5}$ and PM_{10} (Table 4). Moreover, O_3 exposure was associated with an increased excess mortality for all-cause, cardiovascular, and respiratory disease by 6.40% (95% eCI, 6.19–6.59%), 6.71% (95% eCI, 6.44–6.97%), and 7.89% (95% eCI, 7.36–8.35%), respectively. Similarly, our stratified analysis by O_3 exposure levels showed that, overall, excess total and cardiovascular mortality due to PM_{10} and $PM_{2.5}$ was higher in the highly O_3 -exposed population than both the low-exposed population and the whole population, though this trend was not significant for excess mortality due to PM_{10} and excess respiratory mortality (Table 4).

4. DISCUSSION

Our two-stage, multicounty, time-series analysis of 4,276,989 individuals provides new ideas for the synergistic control of multiple air pollutants. Our results indicated that short-term exposure to PM_{10} , $PM_{2.5}$, PM_{10} , and O_3 all increase the risks of daily all-cause, cardiovascular, and respiratory mortality, with associations modified by sex, age, and season. In addition, we found an interaction between PM and O_3 short-term exposure

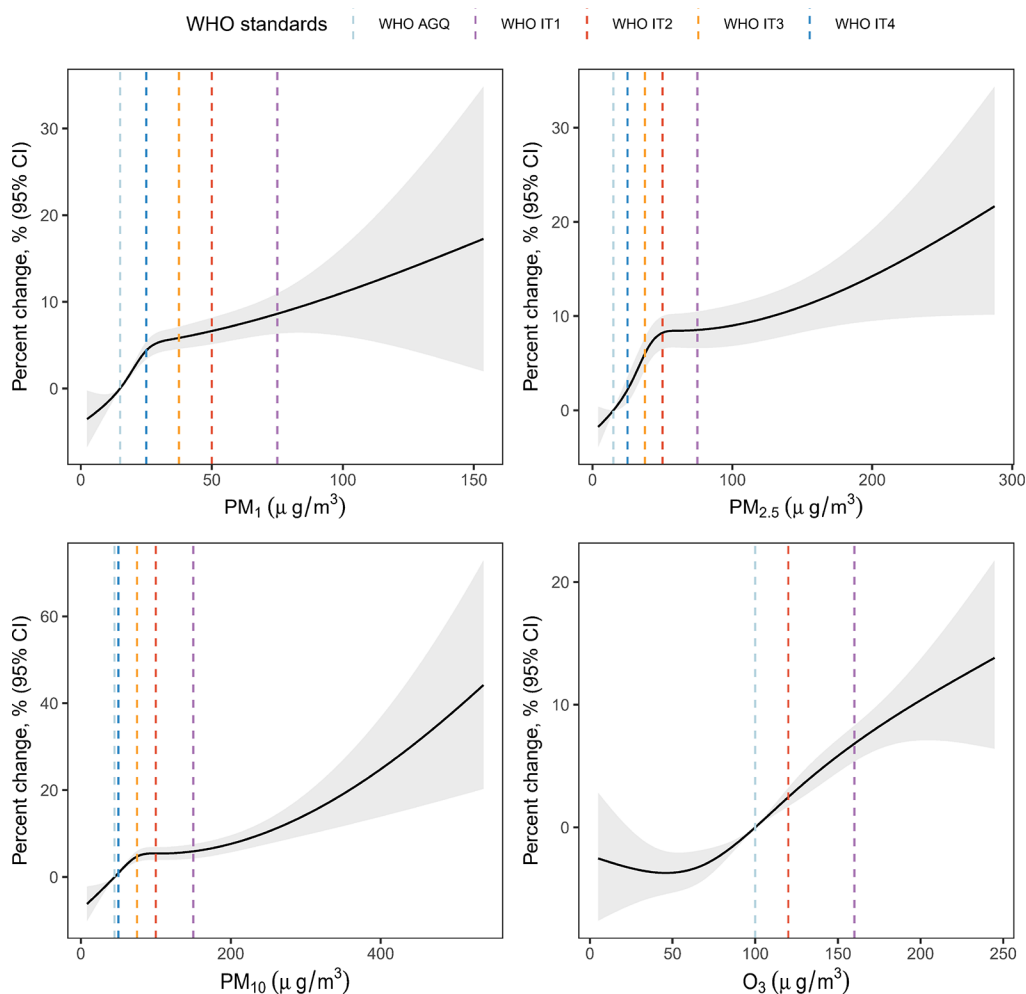


Figure 3. Pooled concentration–response curves for respiratory mortality. Shown are the pooled concentration–response curves for the associations of the 2 day moving average concentrations of PM₁, PM_{2.5}, PM₁₀, and O₃ with daily respiratory mortality. The y-axis displays the percentage difference in mortality from the pooled mean effect. The dashed lines indicate the air quality guidelines or standards for particulate matter (PM₁, PM_{2.5}, or PM₁₀) or ozone (8 h average concentrations) according to the World Health Organization (WHO) Air Quality Guidelines as well as four WHO Interim Targets (IT-1, IT-2, IT-3, and IT-4).

on the daily mortality risk, implying that those with higher short-term exposure levels of both PM and O₃ had the highest risk of mortality. The excess total and cardiovascular mortality due to PM₁ and PM_{2.5} was higher in the highly exposed O₃ population than in the low-exposed population and the population as a whole. This study will provide new insights into the potential mechanisms and synergistic regulatory measures for co-exposure to PM- and O₃-induced mortality and have implications for related policies and standards.

This study found that with a 10 $\mu\text{g}/\text{m}^3$ increase in PM₁, PM_{2.5}, and PM₁₀, the risk of all-cause mortality increased by 1.37%, 0.60%, and 0.36%, respectively. Similar associations were found between PM₁, PM_{2.5}, and PM₁₀ exposure and cardiovascular and respiratory mortality. In line with most previous studies, some studies have found a positive association between short-term exposure to PM₁, PM_{2.5}, and PM₁₀ and daily all-cause mortality, without significant thresholds.^{3,4} Additionally, our study also found a stronger association between pollutants and daily respiratory-related mortality and between the risk of daily mortality and PM₁, which corresponds to the results of other studies.^{4,32–34} However, an epidemiological study based in Barceló, Spain, reported inconsistent results,³⁵ and the association of

pollutants with daily mortality varies from study to study. The inconsistency of these results may be due on one hand to inconsistencies in the sources and components and levels of PM in different regions³⁶ and on the other hand to differences in statistical methods or in the setting of parameterization for the same statistical method.³⁷ Furthermore, we found a 0.80% (95% CI, 0.73–0.87%) increase in all-cause mortality per 10 $\mu\text{g}/\text{m}^3$ increase in O₃. This finding is supported by the most recent meta-analysis, which found a significant positive association between short-term exposure to O₃ and the risk of all-cause mortality (RR, 1.0043; 95% CI, 1.0034–1.0052).⁵ In addition, some studies have found an association between O₃ exposure and death from respiratory disease and cardiovascular disease.^{38,39} These findings all suggest that O₃ is an important contributor to the increased risk of death.

The results of the stratified analysis indicated that short-term exposures to PM and O₃ had a stronger association with daily all-cause and cardiovascular mortality during the warm season, whereas the associations did not differ significantly between the warm and cold seasons for respiratory diseases. This finding is still supported by a number of studies and may be explained by the fact that warmer seasons increase the opportunity for people to travel and thus increase outdoor

Table 3. Multiplicative Interaction between Particulate Matter and Ozone on All-Cause, Cardiovascular, and Respiratory Mortality at the Provincial Level^a

| variables | PM ₁ | | PM _{2.5} | | PM ₁₀ | |
|------------------------------|---------------------------------|----------------------------------|---------------------------------|----------------------------------|---------------------------------|----------------------------------|
| | pooled estimate (%) (95% CI) | multiplicative scale (95% CI) | pooled estimate (%) (95% CI) | multiplicative scale (95% CI) | pooled estimate (%) (95% CI) | multiplicative scale (95% CI) |
| All-Cause Mortality | | | | | | |
| high O ₃ exposure | 2.08 (1.80, 2.36) | 0.14 (0.13, 0.16) | 1.00 (0.85, 1.14) | 0.07 (0.07, 0.08) | 0.31 (0.24, 0.39) | 0.04 (0.04, 0.05) |
| low O ₃ exposure | 0.77 (0.63, 0.91) | | 0.36 (0.29, 0.43) | | 0.23 (0.19, 0.28) | |
| Cardiovascular Mortality | | | | | | |
| high O ₃ exposure | 2.62 (2.16, 3.08) | 0.16 (0.14, 0.17) | 1.28 (1.05, 1.51) | 0.08 (0.07, 0.09) | 0.38 (0.27, 0.49) | 0.04 (0.04, 0.05) |
| low O ₃ exposure | 0.76 (0.56, 0.95) | | 0.36 (0.26, 0.45) | | 0.23 (0.16, 0.30) | |
| Respiratory Mortality | | | | | | |
| high O ₃ exposure | 1.38 (0.59, 2.16) | 0.18 (0.14, 0.21) | 0.64 (0.21, 1.07) | 0.06 (0.05, 0.07) | 0.24 (0.01, 0.48) | 0.09 (0.07, 0.11) |
| low O ₃ exposure | 1.03 (0.62, 1.43) | | 0.48 (0.28, 0.68) | | 0.34 (0.20, 0.48) | |

^aPooled estimates represent the percentage changes in daily all-cause, cardiovascular, and respiratory mortality per 10 $\mu\text{g}/\text{m}^3$ increase in the concentrations of particulate matter (PM) with an aerodynamic diameter of 1 μm or less (PM₁), 2.5 μm or less (PM_{2.5}), and 10 μm or less (PM₁₀) stratified by ozone (O₃). Multiplicative scales represent the percentage changes in the interactions of PM and O₃ on daily all-cause, cardiovascular, and respiratory mortality.

Table 4. Excess Mortality Fraction (%) Attributable to Short-Term Exposure to PM₁, PM_{2.5}, PM₁₀ and O₃^a

| pollutants | excess fraction (%) (95% eCI) | | |
|------------------------------|-------------------------------|--------------------------|-----------------------|
| | all-cause mortality | cardiovascular mortality | respiratory mortality |
| O ₃ | 6.40 (6.19, 6.59) | 6.71 (6.44, 6.97) | 7.89 (7.36, 8.35) |
| PM ₁ | | | |
| high O ₃ exposure | 4.87 (4.62, 5.11) | 6.42 (6.06, 6.74) | 3.91 (3.38, 4.36) |
| low O ₃ exposure | 2.40 (2.35, 2.45) | 2.45 (2.37, 2.53) | 2.94 (2.65, 3.21) |
| all | 3.76 (3.65, 3.86) | 4.06 (3.93, 4.18) | 4.39 (4.01, 4.72) |
| PM _{2.5} | | | |
| high O ₃ exposure | 3.84 (3.60, 4.05) | 5.26 (4.93, 5.54) | 2.92 (2.50, 3.30) |
| low O ₃ exposure | 2.09 (2.04, 2.14) | 2.16 (2.08, 2.24) | 2.66 (2.32, 2.94) |
| all | 2.89 (2.82, 2.96) | 3.17 (3.10, 3.24) | 3.39 (3.05, 3.69) |
| PM ₁₀ | | | |
| high O ₃ exposure | 2.63 (2.42, 2.82) | 3.44 (3.17, 3.68) | 1.96 (1.66, 2.24) |
| low O ₃ exposure | 2.31 (2.25, 2.35) | 2.32 (2.22, 2.40) | 3.08 (2.75, 3.37) |
| all | 3.04 (2.93, 3.14) | 3.23 (3.10, 3.34) | 4.11 (3.73, 4.43) |

^aAbbreviations: PM₁, particulate matter (PM) with an aerodynamic diameter of 1 μm or less; PM_{2.5}, PM with an aerodynamic diameter of 2.5 μm or less; PM₁₀, PM with an aerodynamic diameter of 10 μm or less; and O₃, ozone. The high O₃ exposure group includes the participants with O₃ exposure levels $\geq 100 \mu\text{g}/\text{m}^3$. The low O₃ exposure group includes the participants with O₃ exposure levels $< 100 \mu\text{g}/\text{m}^3$.

exposure levels.⁴⁰ We also found that women and the elderly may be more sensitive to PM and O₃ than men and younger people, which is in line with some previous studies.^{32,41,42} One possible reason is that women have a higher airway responsiveness and have a greater physiological response to air pollution.⁴³ The higher prevalence of cardiovascular disease in the elderly may explain their greater susceptibility to air pollution. There are many other factors that may influence the

associations between air pollution and mortality, such as education and marriage, which is out of the scope of this study and requires further investigation.

Our study found a multiplicative interaction of PM (including PM₁, PM_{2.5}, and PM₁₀) and O₃ on all-cause, cardiovascular, and respiratory mortality, with the strongest interaction between PM₁ and O₃. Most previous studies have either examined the effects of the two pollutants individually or have used dual-pollutant models to explore the modifying effects of O₃ on the association between PM and daily mortality. Our study is supported by a number of previous studies. A study based on a Moscow population found that the relationship between PM₁₀ and mortality was significantly modified by O₃ levels. On days when O₃ concentrations exceeded the 90th percentile, the risk of all-cause mortality was tripled by PM₁₀.⁴⁴ Another study also found an interaction between short-term PM and O₃ exposure on daily mortality.⁴⁵ A similar association was found in several cross-sectional studies and prospective studies.^{46,47} In an animal study, it was observed that treatment with a TRPV1 antagonist reduced both IgE and OVA-IgE levels when co-exposed to PM_{2.5} and O₃. This finding suggests that neuro-immune interactions involving PM_{2.5} and O₃ contribute to the exacerbation of immunoglobulin levels.¹²

Although the mechanisms involved are not fully elucidated, there are currently some possible underlying mechanisms that could be used to suggest that combined exposure to O₃ and PM may have adverse effects. First, a previous study found that participants had a higher susceptibility to death from O₃ in areas with elevated PM_{2.5} oxidation potential and abundant transition metals/sulfur.⁴⁸ Second, PM and O₃ have been found to act synergistically to generate a sustained production of reactive HO radicals, which cause airway inflammation and other respiratory diseases.¹³ In addition, PM_{2.5} and O₃ co-exposure may be involved in the development of adverse outcomes through neuro-immune interactions.¹² Finally, smaller PM sizes (such as PM₁) may be more likely to cause adverse health effects because smaller PM particles are more likely to enter the circulatory system through the respiratory

tract have a larger surface area to adsorb more harmful pollutants.^{49,50}

Our study has the following strengths. First, to the best of our knowledge, our study is the largest and most recent to examine the interaction of short-term co-exposure to PM and O₃ on daily mortality, which will provide a theoretical basis for the combined regulation of both. Furthermore, we explored the susceptibility factors that influence the associations of PM₁, PM_{2.5}, PM₁₀, and O₃ with daily mortality, which will facilitate the provision of individualized protective measures for susceptible populations.

This study also has some limitations. First, exposure misclassification due to our inability to obtain individual exposure data by collecting exposure levels at the county level is an acknowledged inherent limitation of environmental epidemiological studies. Second, the number of deaths from respiratory disease across counties in our study was small, and although we used zero-inflated Poisson regression to improve on this weakness,²⁴ we still could find that the results of subgroup analyses were not very robust. Further, although we controlled for time-varying weather conditions, there were many confounding factors (age, sex, lifestyle, etc.) that could not be controlled, which would lead to imprecise estimates. For this reason, we conducted stratified analyses by sex, age, and season to further examine the modifying effect of these factors on the associations. Moreover, based on previous studies,^{30,41} we calculated the excess mortality fraction attributable to short-term exposure to PM₁, PM_{2.5}, and PM₁₀ by assuming a linear relationship between PM and mortality. However, it is undeniable that this may have led to some misestimation of the excess mortality fraction. The limitations of this indicator should be recognized in practical applications. Finally, although we included all deaths in the Jiangsu mortality surveillance system for the period of 2014–2021, Jiangsu is only one province in China, which limits the generalizability and extrapolation of our results. In the future, accurate exposure measurements and a nationally representative study population will increase the accuracy of such studies and make the results more instructive.

■ ASSOCIATED CONTENT

SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.est.3c04056>.

Summary of air pollution levels, weather conditions, and mortality over the entire study period (Tables S1 and S2 and Figures S1 and S2); the effect estimates and the average scores of GCV for mortality on different lag days of air pollution and temperature (Tables S3 and S4); sensitivity analyses (Tables S5 and S6); stratified analysis by sex, age, and season (Tables S7–S9); and the additive interaction between PM and O₃ on mortality (Figures S3–S5) (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Jing Li – Institute of Child and Adolescent Health, School of Public Health, Peking University, Beijing 100191, China;

● orcid.org/0000-0001-7682-4311; Email: jing.li@hsc.pku.edu.cn

Shaodan Huang – Department of Occupational and Environmental Health Sciences, School of Public Health,

Peking University, Beijing 100191, China; Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing 100191, China;

Email: shuang@bjmu.edu.cn

Hao Yu – Non-Communicable Chronic Disease Control and Prevention Institute, Jiangsu Provincial Center for Disease Control and Prevention, Jiangsu 210009, China; Email: yuh@jscdc.cn

Authors

Jianhui Guo – Institute of Child and Adolescent Health, School of Public Health, Peking University, Beijing 100191, China

Jinyi Zhou – Non-Communicable Chronic Disease Control and Prevention Institute, Jiangsu Provincial Center for Disease Control and Prevention, Jiangsu 210009, China

Renqiang Han – Non-Communicable Chronic Disease Control and Prevention Institute, Jiangsu Provincial Center for Disease Control and Prevention, Jiangsu 210009, China

Yaqi Wang – Institute of Child and Adolescent Health, School of Public Health, Peking University, Beijing 100191, China

Xinyao Lian – Institute of Child and Adolescent Health, School of Public Health, Peking University, Beijing 100191, China

Ziqi Tang – Institute of Child and Adolescent Health, School of Public Health, Peking University, Beijing 100191, China

Jin Ye – School of Energy and Power, Jiangsu University of Science and Technology, Jiangsu 212100, China

Xueqiong He – Department of Occupational and Environmental Health Sciences, School of Public Health, Peking University, Beijing 100191, China

Complete contact information is available at:

<https://pubs.acs.org/10.1021/acs.est.3c04056>

Notes

Consent to participate: Data were analyzed at the aggregate level and no participants were contacted.

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (NSFC) (Grants 42307133, 52208092, and 7219900083), the National Key Research and Development Program of China (Grant 2022YFC3703502), and the Special Fund of the State Key Joint Laboratory of Environmental Simulation and Pollution Control (Grant 23K02ESPCP).

■ REFERENCES

(1) Review of evidence on health aspects of air pollution – REVIHAAP Project: Technical Report; WHO Regional Office for Europe: Copenhagen, 2013.

(2) Fuller, R.; Landrigan, P. J.; Balakrishnan, K.; Bathan, G.; Bose-O'Reilly, S.; Brauer, M.; Caravanos, J.; Chiles, T.; Cohen, A.; Corra, L.; Cropper, M.; Ferraro, G.; Hanna, J.; Hanrahan, D.; Hu, H.; Hunter, D.; Janata, G.; Kupka, R.; Lanphear, B.; Lichtveld, M.; Martin, K.; Mustapha, A.; Sanchez-Triana, E.; Sandilya, K.; Schaeffli, L.; Shaw, J.; Seddon, J.; Suk, W.; Téllez-Rojo, M. M.; Yan, C. Pollution and health: a progress update. *Lancet Planet Health* **2022**, *6* (6), No. e535–e547.

(3) Liu, C.; Chen, R.; Sera, F.; Vicedo-Cabrera, A. M.; Guo, Y.; Tong, S.; Coelho, M.; Saldiva, P. H. N.; Lavigne, E.; Matus, P.; Valdes Ortega, N.; Osorio Garcia, S.; Pascal, M.; Stafoggia, M.; Scortichini, M.; Hashizume, M.; Honda, Y.; Hurtado-Diaz, M.; Cruz, J.; Nunes,

- B.; Teixeira, J. P.; Kim, H.; Tobias, A.; Íñiguez, C.; Forsberg, B.; Åström, C.; Ragettli, M. S.; Guo, Y. L.; Chen, B. Y.; Bell, M. L.; Wright, C. Y.; Scovronick, N.; Garland, R. M.; Milojevic, A.; Kyselý, J.; Urban, A.; Orru, H.; Indermitte, E.; Jaakkola, J. J. K.; Rytli, N. R. L.; Katsouyanni, K.; Analitis, A.; Zanobetti, A.; Schwartz, J.; Chen, J.; Wu, T.; Cohen, A.; Gasparrini, A.; Kan, H. Ambient Particulate Air Pollution and Daily Mortality in 652 Cities. *N Engl J. Med.* **2019**, *381* (8), 705–715.
- (4) Chen, R.; Yin, P.; Meng, X.; Liu, C.; Wang, L.; Xu, X.; Ross, J. A.; Tse, L. A.; Zhao, Z.; Kan, H.; Zhou, M. Fine Particulate Air Pollution and Daily Mortality. A Nationwide Analysis in 272 Chinese Cities. *Am. J. Respir Crit Care Med.* **2017**, *196* (1), 73–81.
- (5) Orellano, P.; Reynoso, J.; Quaranta, N.; Bardach, A.; Ciapponi, A. Short-term exposure to particulate matter (PM(10) and PM(2.5)), nitrogen dioxide (NO(2)), and ozone (O(3)) and all-cause and cause-specific mortality: Systematic review and meta-analysis. *Environ. Int.* **2020**, *142*, 105876.
- (6) Ojha, N.; Soni, M.; Kumar, M.; Gunthe, S. S.; Chen, Y.; Ansari, T. U. Mechanisms and Pathways for Coordinated Control of Fine Particulate Matter and Ozone. *Curr. Pollut Rep* **2022**, *8* (4), 594–604.
- (7) Jin, Y.; Andersson, H.; Zhang, S. Air Pollution Control Policies in China: A Retrospective and Prospects. *Int. J. Environ. Res. Public Health* **2016**, *13* (12), 1219.
- (8) Song, X. H.; Yan, L.; Liu, W.; He, J. Y.; Wang, Y. C.; Huang, T. L.; Li, Y. Y.; Chen, M.; Meng, J. J.; Hou, Z. F. [Spatiotemporal Distribution Characteristics of Co-pollution of PM(2.5) and Ozone over BTH with Surrounding Area from 2015 to 2021]. *Huan Jing Ke Xue* **2023**, *44* (4), 1841–1851.
- (9) Zhang, N.; Guan, Y.; Jiang, Y.; Zhang, X.; Ding, D.; Wang, S. Regional demarcation of synergistic control for PM(2.5) and ozone pollution in China based on long-term and massive data mining. *Sci. Total Environ.* **2022**, *838*, 155975.
- (10) Bates, J. T.; Fang, T.; Verma, V.; Zeng, L.; Weber, R. J.; Tolbert, P. E.; Abrams, J. Y.; Sarnat, S. E.; Klein, M.; Mulholland, J. A.; Russell, A. G. Review of Acellular Assays of Ambient Particulate Matter Oxidative Potential: Methods and Relationships with Composition, Sources, and Health Effects. *Environ. Sci. Technol.* **2019**, *53* (8), 4003–4019.
- (11) Gao, D.; Ripley, S.; Weichenthal, S.; Godri Pollitt, K. J. Ambient particulate matter oxidative potential: Chemical determinants, associated health effects, and strategies for risk management. *Free Radic Biol. Med.* **2020**, *151*, 7–25.
- (12) Lian, Z.; Qi, H.; Liu, X.; Zhang, Y.; Xu, R.; Yang, X.; Zeng, Y.; Li, J. Ambient ozone, and urban PM(2.5) co-exposure, aggravate allergic asthma via transient receptor potential vanilloid 1-mediated neurogenic inflammation. *Ecotoxicol Environ. Saf* **2022**, *243*, 114000.
- (13) Valavanidis, A.; Loidas, S.; Vlahogianni, T.; Fiotakis, K. Influence of ozone on traffic-related particulate matter on the generation of hydroxyl radicals through a heterogeneous synergistic effect. *J. Hazard Mater.* **2009**, *162* (2–3), 886–92.
- (14) Luben, T. J.; Buckley, B. J.; Patel, M. M.; Stevens, T.; Coffman, E.; Rappazzo, K. M.; Owens, E. O.; Hines, E. P.; Moore, D.; Painter, K.; Jones, R.; Datko-Williams, L.; Wilkie, A. A.; Madden, M.; Richmond-Bryant, J. A cross-disciplinary evaluation of evidence for multipollutant effects on cardiovascular disease. *Environ. Res.* **2018**, *161*, 144–152.
- (15) Kalantzi, E. G.; Makris, D.; Duquenne, M. N.; Kaklamani, S.; Stapountzis, H.; Gourgoulis, K. I. Air pollutants and morbidity of cardiopulmonary diseases in a semi-urban Greek peninsula. *Atmospheric environment* **2011**, *45* (39), 7121–7126.
- (16) Fakhri, A. A.; Ilic, L. M.; Wellenius, G. A.; Urch, B.; Silverman, F.; Gold, D. R.; Mittleman, M. A. Autonomic effects of controlled fine particulate exposure in young healthy adults: effect modification by ozone. *Environ. Health Perspect* **2009**, *117* (8), 1287–92.
- (17) Braithwaite, I.; Zhang, S.; Kirkbride, J. B.; Osborn, D. P. J.; Hayes, J. F. Air Pollution (Particulate Matter) Exposure and Associations with Depression, Anxiety, Bipolar, Psychosis and Suicide Risk: A Systematic Review and Meta-Analysis. *Environ. Health Perspect* **2019**, *127* (12), 126002.
- (18) Wei, J.; Li, Z.; Guo, J.; Sun, L.; Huang, W.; Xue, W.; Fan, T.; Cribb, M. Satellite-Derived 1-km-Resolution PM(1) Concentrations from 2014 to 2018 across China. *Environ. Sci. Technol.* **2019**, *53* (22), 13265–13274.
- (19) Wei, J.; Li, Z.; Lyapustin, A.; Sun, L.; Peng, Y.; Xue, W.; Su, T.; Cribb, M. Reconstructing 1-km-resolution high-quality PM2.5 data records from 2000 to 2018 in China: spatiotemporal variations and policy implications. *Remote Sensing of Environment* **2021**, *252*, 112136.
- (20) Wei, J.; Li, Z.; Xue, W.; Sun, L.; Fan, T.; Liu, L.; Su, T.; Cribb, M. The ChinaHighPM(10) dataset: generation, validation, and spatiotemporal variations from 2015 to 2019 across China. *Environ. Int.* **2021**, *146*, 106290.
- (21) Xue, T.; Zheng, Y.; Geng, G.; Xiao, Q.; Meng, X.; Wang, M.; Li, X.; Wu, N.; Zhang, Q.; Zhu, T. Estimating Spatiotemporal Variation in Ambient Ozone Exposure during 2013–2017 Using a Data-Fusion Model. *Environ. Sci. Technol.* **2020**, *54* (23), 14877–14888.
- (22) Muñoz-Sabater, J.; Dutra, E.; Agustí-Panareda, A.; Albergel, C.; Arduini, G.; Balsamo, G.; Boussetta, S.; Choulga, M.; Harrigan, S.; Hersbach, H.; Martens, B.; Miralles, D. G.; Piles, M.; Rodríguez-Fernández, N. J.; Zsoter, E.; Buontempo, C.; Thépaut, J. N. ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. Data* **2021**, *13* (9), 4349–4383.
- (23) Bell, M. L.; Dominici, F.; Samet, J. M. A meta-analysis of time-series studies of ozone and mortality with comparison to the national morbidity, mortality, and air pollution study. *Epidemiology* **2005**, *16* (4), 436–45.
- (24) Long, D. L.; Preisser, J. S.; Herring, A. H.; Golin, C. E. A marginalized zero-inflated Poisson regression model with overall exposure effects. *Stat Med.* **2014**, *33* (29), 5151–65.
- (25) Golub, G. H.; Heath, M.; Wahba, G. Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics* **1979**, *21* (2), 215–223.
- (26) Peng, R. D.; Chang, H. H.; Bell, M. L.; McDermott, A.; Zeger, S. L.; Samet, J. M.; Dominici, F. Coarse particulate matter air pollution and hospital admissions for cardiovascular and respiratory diseases among Medicare patients. *Jama* **2008**, *299* (18), 2172–9.
- (27) Samet, J. M.; Dominici, F.; Curriero, F. C.; Coursac, I.; Zeger, S. L. Fine particulate air pollution and mortality in 20 U.S. cities, 1987–1994. *N Engl J. Med.* **2000**, *343* (24), 1742–9.
- (28) Atkinson, R. W.; Kang, S.; Anderson, H. R.; Mills, I. C.; Walton, H. A. Epidemiological time series studies of PM2.5 and daily mortality and hospital admissions: a systematic review and meta-analysis. *Thorax* **2014**, *69* (7), 660–5.
- (29) Gasparrini, A.; Leone, M. Attributable risk from distributed lag models. *BMC Med. Res. Methodol* **2014**, *14*, 55.
- (30) O'Brien, E.; Masselot, P.; Sera, F.; Roye, D.; Breitner, S.; Ng, C. F. S.; de Sousa Zanotti Stagliorio Coelho, M.; Madureira, J.; Tobias, A.; Vicedo-Cabrera, A. M.; Bell, M. L.; Lavigne, E.; Kan, H.; Gasparrini, A. Short-Term Association between Sulfur Dioxide and Mortality: A Multicountry Analysis in 399 Cities. *Environ. Health Perspect* **2023**, *131* (3), 37002.
- (31) Di, Q.; Dai, L.; Wang, Y.; Zanobetti, A.; Choirat, C.; Schwartz, J. D.; Dominici, F. Association of Short-term Exposure to Air Pollution With Mortality in Older Adults. *Jama* **2017**, *318* (24), 2446–2456.
- (32) Hu, K.; Guo, Y.; Hu, D.; Du, R.; Yang, X.; Zhong, J.; Fei, F.; Chen, F.; Chen, G.; Zhao, Q.; Yang, J.; Zhang, Y.; Chen, Q.; Ye, T.; Li, S.; Qi, J. Mortality burden attributable to PM(1) in Zhejiang province, China. *Environ. Int.* **2018**, *121*, 515–522.
- (33) Lin, H.; Tao, J.; Du, Y.; Liu, T.; Qian, Z.; Tian, L.; Di, Q.; Rutherford, S.; Guo, L.; Zeng, W.; Xiao, J.; Li, X.; He, Z.; Xu, Y.; Ma, W. Particle size and chemical constituents of ambient particulate pollution associated with cardiovascular mortality in Guangzhou, China. *Environ. Pollut.* **2016**, *208*, 758–766.
- (34) Lin, H.; Tao, J.; Du, Y.; Liu, T.; Qian, Z.; Tian, L.; Di, Q.; Zeng, W.; Xiao, J.; Guo, L.; Li, X.; Xu, Y.; Ma, W. Differentiating the effects of characteristics of PM pollution on mortality from ischemic and hemorrhagic strokes. *Int. J. Hyg Environ. Health* **2016**, *219* (2), 204–11.

(35) Xiong, J.; Li, J.; Wu, X.; Wolfson, J. M.; Lawrence, J.; Stern, R. A.; Koutrakis, P.; Wei, J.; Huang, S. The association between daily-diagnosed COVID-19 morbidity and short-term exposure to PM(1) is larger than associations with PM(2.5) and PM(10). *Environ. Res.* **2022**, *210*, 113016.

(36) Shi, Z.; Li, J.; Huang, L.; Wang, P.; Wu, L.; Ying, Q.; Zhang, H.; Lu, L.; Liu, X.; Liao, H.; Hu, J. Source apportionment of fine particulate matter in China in 2013 using a source-oriented chemical transport model. *Sci. Total Environ.* **2017**, *601–602*, 1476–1487.

(37) Pui, D. Y. H.; Chen, S.-C.; Zuo, Z. PM2.5 in China: Measurements, sources, visibility and health effects, and mitigation. *Particulology* **2014**, *13*, 1–26.

(38) Liu, W.; Wei, J.; Cai, M.; Qian, Z.; Long, Z.; Wang, L.; Vaughn, M. G.; Aaron, H. E.; Tong, X.; Li, Y.; Yin, P.; Lin, H.; Zhou, M. Particulate matter pollution and asthma mortality in China: A nationwide time-stratified case-crossover study from 2015 to 2020. *Chemosphere* **2022**, *308*, 136316.

(39) Xu, R.; Wang, Q.; Wei, J.; Lu, W.; Wang, R.; Liu, T.; Wang, Y.; Fan, Z.; Li, Y.; Xu, L.; Shi, C.; Li, G.; Chen, G.; Zhang, L.; Zhou, Y.; Liu, Y.; Sun, H. Association of short-term exposure to ambient air pollution with mortality from ischemic and hemorrhagic stroke. *Eur. J. Neurol* **2022**, *29* (7), 1994–2005.

(40) Calkins, M.; Szmerkovsky, J. G.; Biddle, S. Effect of Increased Time Spent Outdoors on Individuals with Dementia Residing in Nursing Homes. *Journal of Housing For the Elderly* **2007**, *21* (3–4), 211–228.

(41) Xu, R.; Wei, J.; Liu, T.; Li, Y.; Yang, C.; Shi, C.; Chen, G.; Zhou, Y.; Sun, H.; Liu, Y. Association of short-term exposure to ambient PM(1) with total and cause-specific cardiovascular disease mortality. *Environ. Int.* **2022**, *169*, 107519.

(42) Shin, H. H.; Gogna, P.; Maquiling, A.; Parajuli, R. P.; Haque, L.; Burr, B. Comparison of hospitalization and mortality associated with short-term exposure to ambient ozone and PM(2.5) in Canada. *Chemosphere* **2021**, *265*, 128683.

(43) Clougherty, J. E. A growing role for gender analysis in air pollution epidemiology. *Environ. Health Perspect* **2010**, *118* (2), 167–76.

(44) Revich, B.; Shaposhnikov, D. The effects of particulate and ozone pollution on mortality in Moscow, Russia. *Air Qual Atmos Health* **2010**, *3* (2), 117–123.

(45) Chen, G. H.; Song, G. X.; Jiang, L. L.; Zhang, Y. H.; Zhao, N. Q.; Chen, B. H.; Kan, H. D. Interaction between ambient particles and ozone and its effect on daily mortality. *Biomed Environ. Sci.* **2007**, *20* (6), 502–5.

(46) Ruan, Z.; Qian, Z. M.; Guo, Y.; Zhou, J.; Yang, Y.; Acharya, B. K.; Guo, S.; Zheng, Y.; Cummings-Vaughn, L. A.; Rigdon, S. E.; Vaughn, M. G.; Chen, X.; Wu, F.; Lin, H. Ambient fine particulate matter and ozone higher than certain thresholds associated with myopia in the elderly aged 50 years and above. *Environ. Res.* **2019**, *177*, 108581.

(47) Siddika, N.; Rantala, A. K.; Antikainen, H.; Balogun, H.; Amegah, A. K.; Rytö, N. R. I.; Kukkonen, J.; Sofiev, M.; Jaakkola, M. S.; Jaakkola, J. J. K. Synergistic effects of prenatal exposure to fine particulate matter (PM(2.5)) and ozone (O(3)) on the risk of preterm birth: A population-based cohort study. *Environ. Res.* **2019**, *176*, 108549.

(48) Toyib, O.; Lavigne, E.; Traub, A.; Umbrio, D.; You, H.; Ripley, S.; Pollitt, K.; Shin, T.; Kulka, R.; Jessiman, B.; Tjepkema, M.; Martin, R.; Stieb, D. M.; Hatzopoulou, M.; Evans, G.; Burnett, R. T.; Weichenthal, S. Long-term Exposure to Oxidant Gases and Mortality: Effect Modification by PM 2.5 Transition Metals and Oxidative Potential. *Epidemiology* **2022**, *33* (6), 767–776.

(49) Kan, H. The smaller, the worse? *Lancet Planet Health* **2017**, *1* (6), e210–e211.

(50) Valavanidis, A.; Fiotakis, K.; Vlachogianni, T. Airborne particulate matter and human health: toxicological assessment and importance of size and composition of particles for oxidative damage and carcinogenic mechanisms. *J. Environ. Sci. Health C Environ. Carcinog Ecotoxicol Rev.* **2008**, *26* (4), 339–62.



Blood multiple heavy metals exposure and lung function in young adults: A prospective Cohort study in China

Minghao Wang^{a,1}, Lailai Yan^{b,c,d,1}, Siqi Dou^{a,1}, Liu Yang^a, Yiwen Zhang^e, Wenzhong Huang^e, Shanshan Li^e, Peng Lu^{a,*}, Yuming Guo^{a,e,**}

^a Binzhou Medical University, Yantai, Shandong, China

^b Department of Laboratorial Science and Technology, School of Public Health, Peking University, Beijing 100191, P. R. China

^c Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education

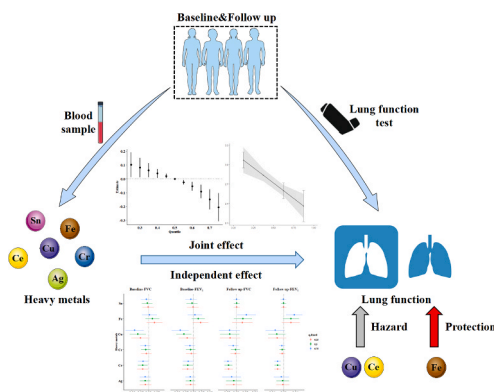
^d Department of Laboratorial Science and Technology & Vaccine Research Center, School of Public Health, Peking University, Beijing 100191, P. R. China

^e Climate, Air Quality Research Unit, School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia

HIGHLIGHTS

- Combined exposure of multiple heavy metals had an adverse effect on lung function.
- Ce and Cu in blood had adverse effects on lung function.
- Fe in blood was a protective factor for lung function.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Lingxin Chen

Keywords:

Heavy metal
Joint effect
Lung function
Cohort study

ABSTRACT

The content of single heavy metal in blood is associated with lung function decline, but there is little evidence on the joint effect of multiple heavy metals on lung function. To explore whether heavy metal mixture exposure is associated with lung function reduction among young adults. The study based on a cohort of 518 students recruited from a college in Shandong, China. We measured their lung function and blood heavy metal concentrations. The BKMR model was used to analyse the association between blood heavy metals mixture levels and lung function, and to identify the critical single heavy metal which contributes most to joint effects. As the sensitivity analysis, we used quantile g-computation model and GLM to explore the joint effect and independent effects of heavy metals. Our findings revealed a significant reduction of FVC and FEV₁ levels after exposure to heavy metals mixture. An IQR increase in Cu was associated with a 0.079 L and 0.083 L decrease in FEV₁ and

* Corresponding author.

** Corresponding author at: Binzhou Medical University, Yantai, Shandong, China.

E-mail addresses: Peng.lu@monash.edu (P. Lu), Yuming.Guo@monash.edu (Y. Guo).

¹ Co-first author of this paper

<https://doi.org/10.1016/j.jhazmat.2023.132064>

Received 6 May 2023; Received in revised form 5 July 2023; Accepted 12 July 2023

Available online 17 July 2023

0304-3894/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).

FVC, respectively. And an IQR increase in Fe was associated with 0.036 L higher FEV₁ and 0.033 L higher FVC. For adults, reducing blood heavy metals concentration might be an effective intervention to protect lung function.

1. Introduction

Heavy metals are defined as metals with relatively high densities, atomic numbers or atomic weights [1]. Due to the rapidly increasing usages in recent years, the pollution caused by heavy metals to the environment has become a great concern [2]. For example, the concentrations of Cadmium(Cd), Nickel(Ni), and Chromium(Cr) in 2012 were 10, 13 and 16 times higher than those in 1941 in Spain, respectively [3]. The concentration of heavy metals in human body could accumulate through a variety of ways like air, foods, drinking water, and skin contact [4,5]. Exposure to heavy metals has been shown to be adversely associated with a series of diseases including cardiovascular diseases [6], reproductive dysfunction [7] and neurological diseases [8]. Mechanism studies have found that heavy metals exposure could increase oxidative stress and inflammatory reaction, which suggest that heavy metals might also cause damage to the respiratory system [9]. Lung function decline is a preclinical symptom of many respiratory diseases [10]. However, evidence on heavy metal exposure and lung function is scarce, especially in low- and middle-income developing countries.

A study among children aged 6–17 years across the US has found a negative association between blood manganese (Mn) and forced vital capacity (FVC), as well as a negative correlation between urine Pb and forced expiratory flow at 25–75%(FEF_{25–75}) [11]. Similar negative associations have also been found in other three studies between cadmium (Cd) [12], Zn [13], and copper (Cu) [14] and lung function. However, these studies only considered the independent effect of single heavy metal on lung function, which did not account for the joint effects of heavy metal mixtures. A cross-sectional study among 186 welders has found a negative joint effect of urine metal mixtures on lung function (FVC, FEV₁, and PEF) [15]. Similar evidence was found in a cross-sectional study among American Indian adults from Strong Heart Study [16]. However, urine metal concentrations used in these studies to assess the heavy metal exposure could be highly affected by individual metabolic abilities compared to the blood metal concentrations. Previous studies have found that blood metal concentrations are closely associated with respiratory diseases [11,17]. Additionally, no studies have been conducted in Chinese general population who lived in heavy metals highly polluted areas [18,19].

We conducted a longitudinal cohort to analyze the association between metal mixtures and lung function in Shandong, China. We aimed to investigate the individual and joint associations between heavy metals in blood and lung function, and to identify the critical metals that contribute the association.

2. Method

2.1. Study design and participants

Participants in the study were from the Chinese Undergraduates Cohort (CUC), a prospective cohort which was described in detail in previous studies [20,21]. A total of 518 participants were recruited in 2019. The inclusion criteria are: (1) Enrolled in Binzhou Medical University in 2019; (2) No hearing, vision or language disability; (3) Will leave to study somewhere else after living in school for a year and a half. We collected their blood samples for further test of heavy metal concentrations, interviewed all participants to complete a questionnaire and examined lung function in baseline survey (September 3 to October 17, 2019). Among all participants, 504 of them finished the follow-up study during May 25 and 26, 2021. The study obtained written informed

consent from all participants. The study was approved by the Binzhou Medical University ethics committee (NO.2019075).

2.2. Lung function test

FVC and forced expiratory volume in 1 s (FEV₁) were selected as indicators of lung function. All the lung function tests were conducted using Gest HI-101 spirometer (Chest, Tokyo, Japan) with a standard procedure according to European Respiratory Society specifications [22]. All spirometers were calibrated before the test. Professionally trained investigators would guide the participant during the test process. All participants were asked to rest few minutes before testing to eliminate potential effect of short-time movement on lung function results, and then each participant completed 3 projects including Slow Vital Capacity (SVC), FVC, and Maximum Ventilatory Volume (MVV), respectively. And all participants were asked to take a rest period of ≥ 1 min between two projects to achieve the best possible result [22]. To ensure the accuracy, we checked the spirometer's diagnosis of results and performed second lung function tests when results suggested pulmonary dysfunction after 20 days of the first test. The same process continued until the third test which was 15 days after the second test. The best result of the three measurements was used to represent the lung function status for participants who were still diagnosed with pulmonary dysfunction at the third time. To avoid the influence of temperature on lung function during the measurement, we used air conditioning to control the temperature in the test room. The flow-chart of the lung function test is shown in Fig. 1.

2.3. Blood sample preparation and metal analysis

Fasting blood sample (approximately 5 mL) was collected before 8:00 a.m. from each participant at both baseline (September 5, 2019) and follow-up (May 27, 2021). Blood samples were centrifuged and divided at room temperature and stored at -80°C prior to the metal analysis.

A direct dilution method was used for the metal measurement. 0.35ml of blood sample were transferred to a quartz tube and combined with 0.40 mL of nitric acid. After predigestion at room temperature for two hours, we placed the quartz tubes in a microwave digestion system (Ultra WAVE, Milestone Co., Italy) for 50 min, and then added 0.1 mL indium (2 ng/mL) as an internal standard element. The concentrations of elements were measured by an inductively coupled plasma-mass spectrometer (ICP-MS, ELAN DRC II, PerkinElmer, USA). There are three main approaches to address the problem when measurements are below the limit of detection (LOD) [23]: A) Removal of the participants whose blood heavy metal levels are below the LOD; B) Replaced it with a random number below the LOD; C) Replaced it with the LOD divided by the square root of 2. Excluding participants would result in missing sample sizes, and using random numbers would increase the heterogeneity. So, we chose to replace measurements below the detection limit with the detection limit divided by the square root of 2 [24].

2.4. Covariates

An electronic questionnaire would be sent to participants by the investigator. If there was any omission or logical error, the investigator would immediately supplement data and provide feedback. Covariates include the demographic information such as sex (male, female, categorical), age (continuous), smoking (never smoking, past smoking, and current smoking, categorical) and alcohol consumption (current

drinking: participants who drank more than once a month on average during the last 12 months; past drinking; never drinking, categorical); and the socioeconomic factor: annual household income ($\leq 20,000$ USD/year, $> 20,000$ USD/year, categorical). The body mass index (BMI, kg/m^2) was calculated by dividing weight in kilograms by the square of height in meters, and controlled as a continuous variable.

2.5. Statistical analysis

We described the demographic characteristics of all participants, where continuous variables were presented as mean \pm standard deviation, while the categorical variables were presented as counts (percentage). Wilcoxon test and Chi-square test were performed for the distribution non-normal continuous and categorical variables between baseline and the follow-up phase.

We included all heavy metals and screened them using a Bayesian kernel machine regression (BKMR) model based on baseline stage. Firstly, we excluded heavy metals that below the detection limit of the spectrometer, the concentration of these heavy metals were lower than the detection accuracy of the spectrometer which might result in inaccurate measurement. Then, we excluded heavy metals with inconsistent association (changed from positive association to negative association) when other heavy metals were controlled at different percentile. Finally, six heavy metals were included in this study, with Ce (Cerium, density: $6.77 \text{ g}/\text{cm}^3$), Ag (Argentum, density: $10.49 \text{ g}/\text{cm}^3$), Cu (Cuprum, density: $8.96 \text{ g}/\text{cm}^3$), Sn (Stannum, density: $7.28 \text{ g}/\text{cm}^3$), Cr (Chromium, density: $7.19 \text{ g}/\text{cm}^3$) and Fe (Ferrum, density: $7.86 \text{ g}/\text{cm}^3$) as the main heavy metals (Fig. S1.).

BKMR was used to model the association between heavy metals and FVC or FEV_1 [25,26]. The BKMR model allows for identifying not only independent effects but also the joint effects via a kernel function. Before modeling, the exposure levels were normalized. In short, we subtracted the mean of the concentration matrix and divided by the standard deviation of the concentration matrix. We created the BKMR model based on the following equation for heavy metals determined in blood samples:

$$Y_i = h(Z) + \beta z_i + e_i$$

Where Y_i was the outcome of lung function (i.e., FVC or FEV_1); $h()$ was the function of fitting exposure and Y_i , which considered both non-linear relationships and joint effect among mixed exposures; Z was the concentrations of heavy metals; z_i was a vector of covariates, including sex, age, BMI, income, smoking and drinking; e_i was a random error term. A "gaussian" link function was used for the BKMR model. The Markov

chain Monte Carlo algorithm of the BKMR model realized 30,000 iterations to ensure the stability of the model results.

As sensitivity analysis, first, we used quantile g-computation (qgcomp) to estimate the association between heavy metals mixture and lung function. It estimates the joint effect of increasing all heavy metals within the mixture by a single quantile. Meanwhile, qgcomp also could obtain the weights of each heavy metal in mixture. Then, we applied Generalized Linear Model (GLM) to evaluate the association between every single metal in heavy metal mixtures and lung function indexes, which was modeled separately between baseline and follow-up. To avoid the difference caused by the distinct order of the magnitude for each heavy metal, we used inter quartile range (IQR) to evaluate the changed effects estimates between pollutants and lung function.

All data analyzed were implemented using R (version 4.1.1) and $p < 0.05$ was considered statistically significant unless otherwise indicated.

3. Result

3.1. Characteristics of the study population

Table 1 shows the demographic characteristics and lung function test results of all participants. Our baseline set included 518 participants,

Table 1
Demographic characteristics and lung function outcomes of all participants.

| Characteristics | Baseline (N = 518) | Follow-up (N = 504) | P-value |
|--------------------------------|--------------------|---------------------|---------|
| Sex | | | 0.86 |
| Male | 162(31.27%) | 154(30.56%) | |
| Female | 356(68.73%) | 350(69.44%) | |
| Age (years) | 18.12 \pm 0.62 | 20.20 \pm 0.67 | < 0.01 |
| BMI (kg/m^2) | 21.86 \pm 3.64 | 22.32 \pm 5.08 | 0.09 |
| Income (USD / year) | | | 0.49 |
| $\leq 20,000$ | 469(90.54%) | 446(88.49%) | |
| $> 20,000$ | 49(9.46%) | 58(11.51%) | |
| Smoking | | | 0.86 |
| Never smoking | 514(99.23%) | 499(99.01%) | |
| Past smoking | 1(0.19%) | 1(0.20%) | |
| Current smoking | 3(0.58%) | 4(0.79%) | |
| Drinking | | | 0.67 |
| Never drinking | 500(96.53%) | 492(97.62%) | |
| Past drinking | 1(0.19%) | 1(0.20%) | |
| Current drinking | 17(3.28%) | 11(2.18%) | |
| Lung function index | | | |
| FVC (L) | 2.72 \pm 0.81 | 3.30 \pm 0.72 | < 0.01 |
| FEV_1 (L) | 2.65 \pm 0.77 | 3.08 \pm 0.63 | < 0.01 |

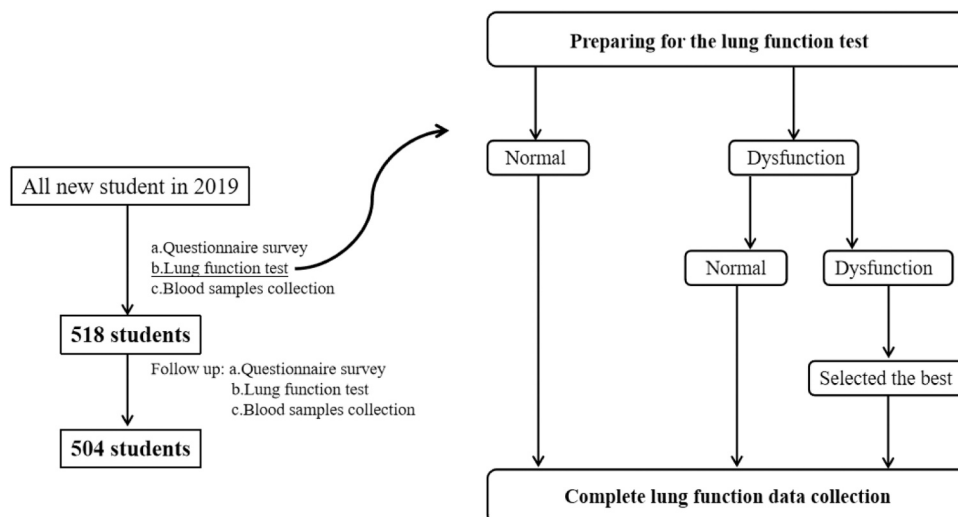


Fig. 1. Flow-chart of study population selection and lung function test.

among whom 356(68.73%) were female. In the follow-up set there were 69.44% female among 504 participants. The average age (\pm SD) of the baseline was 18.12 ± 0.62 years old. The average (\pm SD) BMI at baseline and follow-up were 21.86 ± 3.64 and 22.32 ± 5.08 kg/m², respectively. The vast majority of the participants had never been smoking or drinking. The baseline and follow-up phase exhibited no statistical differences in BMI, annual income, nor in smoking and drinking status. We controlled age in further analysis to eliminate the potential confounding effect. Fig. S2. shows the geographical distribution of participants before baseline survey.

3.2. Blood metal levels

Among 6 heavy metals in this study, the top three most abundant heavy metals were Fe, Cu and Cr on average, and the concentration of other three heavy metals are relatively low. In the baseline survey, the average concentrations of Fe, Cu and Cr in the blood of participants reached 495585.97 ng/mL, 838.63 ng/mL and 3.00 ng/mL. The concentrations of these three heavy metals decreased at different levels during follow up, among which iron and copper decreased significantly. In addition, Ag also decreased significantly, while Sn and Ce increased significantly. The concentration of heavy metals in baseline and follow up was list in Table S1.

3.3. Joint effects of the mixtures of metals on FVC and FEV₁

The joint effects of heavy metal exposure on FVC and FEV₁ (estimates and 95% credible intervals, CIs) are summarized in Fig. 2. The graph represents the estimated changes in lung function associated with percentile changes of heavy metal mixtures, with the 50th percentile of multiple exposure as the reference. From the results of BKMR, as the concentration of heavy metals increases, the joint effect of heavy metals will reduce FVC and FEV₁. The results indicated a negative association

between the whole metal mixture and lung function in baseline and follow-up. The strongest negative association was found in participants whose heavy metal mixtures concentrations were at high percentiles (i. e., > 50th).

3.4. Critical metals and independent effect

To further explore the contribution of critical heavy metal to the joint effects, this study analyzed the independent effect of single metal on FVC and FEV₁. Fig. 3. shows the effect of these metals on lung function indexes when the concentration of other metals was fixed at the 75th, 50th or 25th percentile. We found that Cu and Ce would significantly defect lung function in the baseline, and the results of follow-up were similar. For instance, FVC were estimated to decrease -0.05 L (95%CI: $-0.10, 0.006$), -0.06 L (95%CI: $-0.11, -0.02$), -0.08 L (95%CI: $-0.14, -0.03$) when Cu change from its 25th percentile to 50th percentile when all of the other heavy metals fixed at the 25th, 50th, 75th percentile respectively in baseline. In addition to the above results, we also found that Fe could significantly improve lung function when other heavy metals were controlled at 25th at the follow-up stage. Meanwhile, we also found that when the percentile of other heavy metals changed (25th, 50th, 75th), the effect of Ce on FVC or FEV₁ basically remain stable. But the effect of Fe and Cu were different when other heavy metals fixed at different percentile, indicating the potential interaction between Fe, Cu and other heavy metals. Posterior inclusion probabilities (PIPs) of all heavy metals were shown in Table S2. PIP of Cu, Ce and Fe were relatively high.

3.5. Sensitivity analyses

From the qgcomp model, every IQR increase of heavy metals mixture at baseline and follow-up resulted in a different degree of reduction in lung function. Fig. 4. shows the reduction of each indicator at baseline

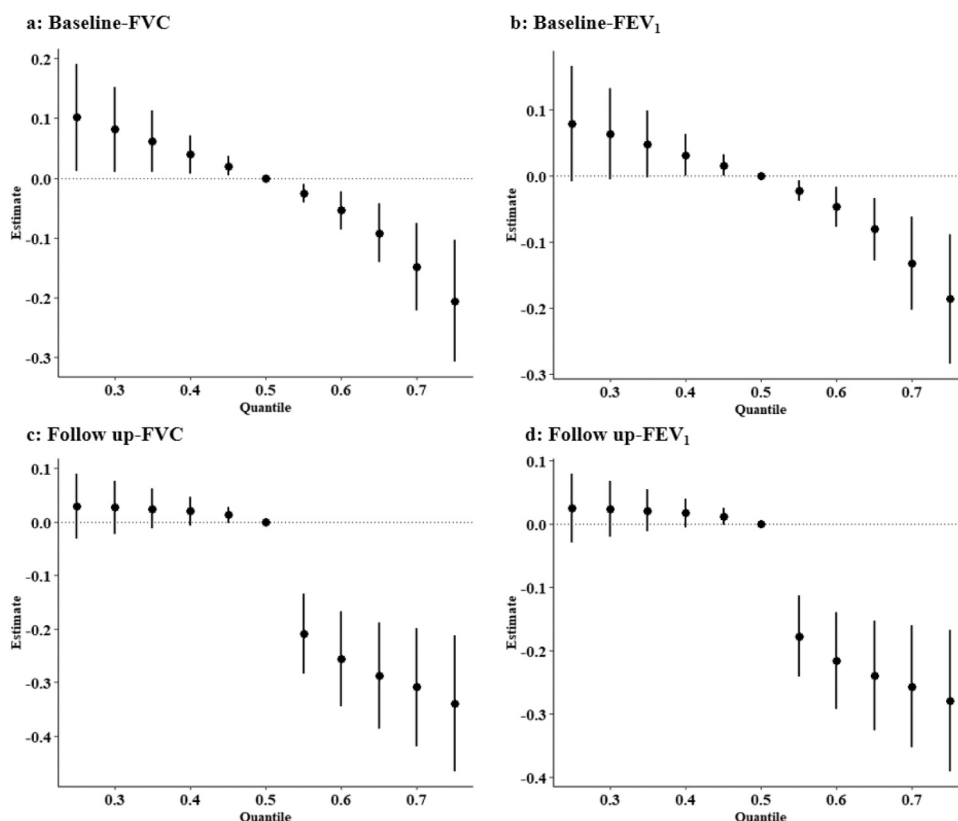


Fig. 2. Joint effects of heavy metals on FVC and FEV₁ determined by the BKMR model with the data from different phase.

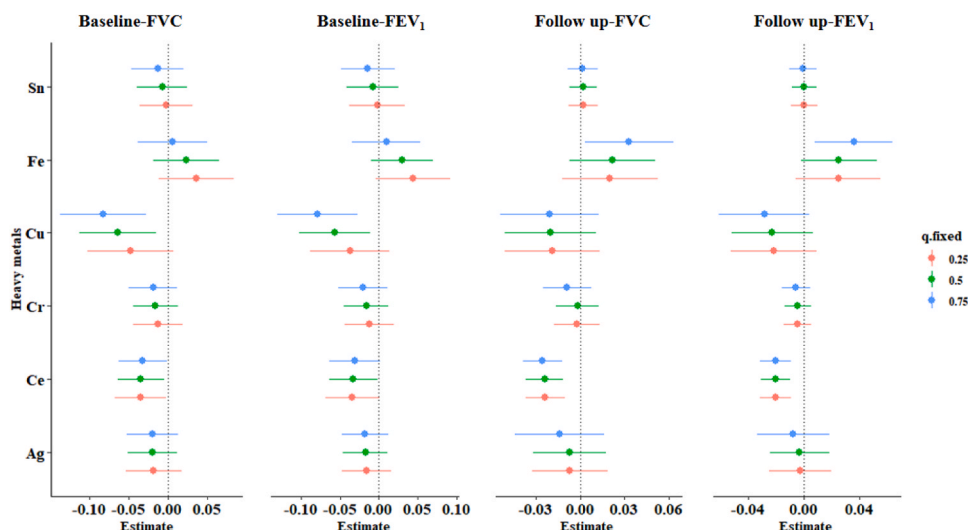


Fig. 3. The estimated change of FVC or FEV₁ associated with the change in a single heavy metal from its 25th percentile to 75th percentile, where all of the other heavy metals are fixed at a particular threshold (25th, 50th, or 75th percentile).

and follow-up stage. The weights for each pollutant were shown in Fig. S3. Table S3 shows the results of GLM in sensitivity analyses. In baseline stage, for every IQR increase of Cu concentration, FVC decreased by 0.1206 (−0.2067, −0.0344) L and FEV₁ decreased by 0.1033 (−0.1895, −0.0344) L, significantly. Ce tended to have a reduction effects on FVC (−0.0072 L, 95CI: −0.0227, 0.0083) and FEV₁ (−0.0054 L, 95CI: −0.0200, 0.0093). Fe had a certain insignificant protective effect (FVC: 0.2206 L, 95CI: −0.6619, 0.9928; FEV₁: 0.2206 L, 95CI: −0.4413, 0.9928). At the follow-up, the significant association between Ce, Cu and Fe and lung function showed a consistent trend with the results of BKMR (Fig. 3.).

4. Discussion

We conducted a longitudinal cohort to analyze the association between metal mixtures and lung function in Shandong, China. Our study found that the combined exposure of multiple heavy metals had an adverse effect on lung function. We identified key metals, e.g. Cu and Ce, which had adverse effects on lung function. In addition, we found that Fe was a protective factor for lung function. This study provided new evidence regarding the effect of mixed heavy metals exposure on lung function, preclinical symptom of many respiratory diseases.

Our study found that the joint effect of multiple heavy metals in blood significantly reduced lung function, both FVC and FEV₁. Studies supporting our findings emerged recently. A cross-sectional study of 186 welders in Anhui province, China found a significant association

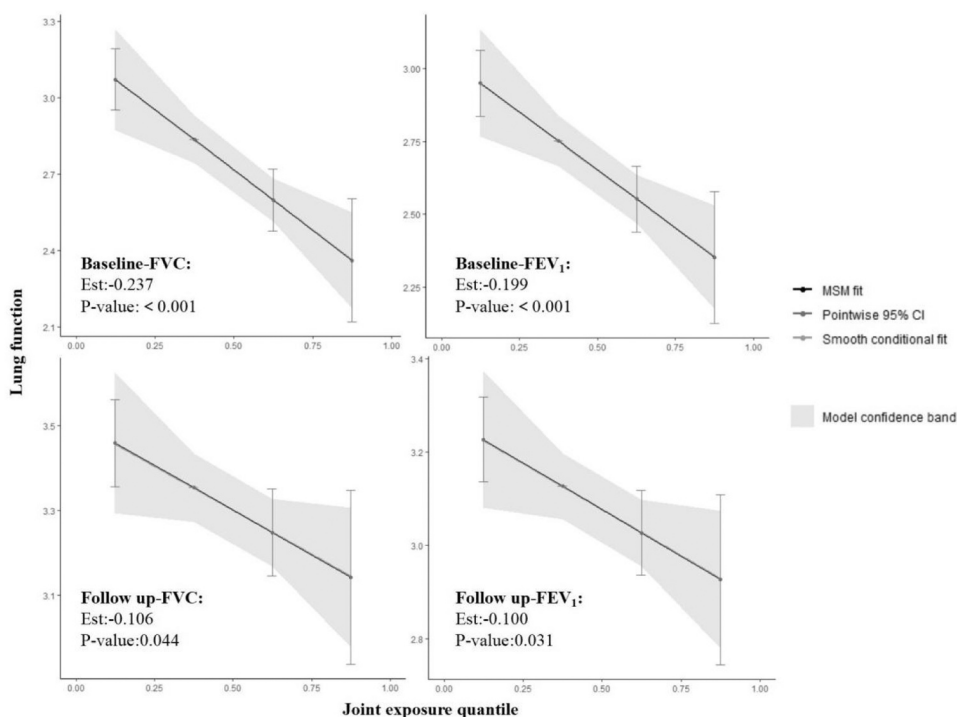


Fig. 4. Joint effects of heavy metals on FVC and FEV₁ determined by the qqcomp model at baseline and follow-up.

between 23 metal mixtures exposure and lung function reduction [15]. Our study found a stronger joint effect of multiple heavy metals and a narrower confidence interval range than the aforementioned studies. Those differences might be explained by the larger sample size in our study. Besides, another cohort study with 2077 participants based on the data of strong heart study cohort in the United States also found that the joint negative effect of multiple metals including As, Cd, Molybdenum (Mo), Selenium (Se), Tungsten(W), and Zn, on FVC [16]. Generally, the results supported that multiple metal exposure impaired lung function. Our results also indicated that the higher concentration of heavy metals, the greater joint effects on lung function. Related biological mechanism studies also supported our finding. A rat-based animal study found that multiple metals can accumulate in the lung and cause damage to lung function [27]. In addition, multiple heavy metals exposure have been confirmed to be associated with inflammation [28] and oxidative stress [29], both of which may negatively affect lung function [28,30,31].

Some critical components, such as Cu and Ce, reduced lung function significantly. Excessive Cu was proved to be associated with respiratory diseases such as asthma [32], COPD [33] and chronic respiratory disease [34]. In a case-control study focusing on COPD, it was found that serum Cu levels were higher in COPD patients, and Cu was negatively associated with FEV₁ and FVC. The study also found changes in biomarkers such as C-reactive protein levels, white blood cell counts, and sedimentation rate, suggesting that the association between Cu and lung function may be mediated by inflammatory-like substances [35]. However, few studies focused on the association between Ce in blood and lung function. Inhalable Ce and CeO₂ can settle in the alveolar area due to small particle size, and could increase lung burden when the deposition concentration exceeds the clearance concentration [36]. Specifically, excessive dose of CeO₂ will lead to proliferation of lung and lymph tissues [36], and inhalation of CeO₂ induces pulmonary inflammation [37]. More studies are needed to further explore the detrimental effects of Ce on lung function.

We found that Fe plays a protective effect on lung function. Fe plays a critical role in many pathways related to the respiratory system, such as oxygen transport, cell respiration, the activity of numerous enzymes, and immune function [38]. However, the effect of Fe on lung function is still unclear. In consistent with our results, a study from the fourth and fifth Korean National Health and Nutrition Examination Survey explored the association between serum Fe and lung function, and it was found that Fe was positively correlated with FEV₁ in the Korean adult population [39]. There are other studies found that Fe deficiency may lead to harmful consequences of respiratory system, such as lung inflammation [40] and asthma [41]. However, another study focusing on participants with COPD found that exposure to Fe in the environment decreased FEV₁ [42]. The difference between these two studies might be due to diverse effects of endogenous and exogenous Fe on human. Meanwhile, the effect of Fe on lung function shows different estimates and significance when the concentration of other heavy metals changes, which indicated that there was a complex interaction between Fe and other heavy metals. Further explorations are needed.

The present study has several strengths. First, our study considered both single and the joint effects of multiple metals on lung function. Multiple exposures can better reflect the real-life exposure conditions than single exposure. Second, we used blood heavy metals concentrations, which was more accurate than external exposure. Third, the BKMR model considered non-linear dose-response and multi-mixture interactions [25]. Moreover, the cohort design could ensure the reliability of our study.

There are also several limitations. First, other heavy metals present in blood that were not considered in this study, and these heavy metals may also have an effect on lung function. However, too much metals in the mixed exposure model might weaken the effect of critical components [25]. Second, this study did not provide experimental validations. More experimental studies are needed to confirm our findings. Finally, in chronological order, the heavy metals in our study were measured

after the lung function test and this might cause certain limitations in causal inference. However, unless a major heavy metal exposure event occurs, the concentration of heavy metals in the human body should not change largely in a short time.

5. Conclusion

Blood heavy metal mixtures could reduce lung function regarding both FVC and FEV₁. The critical heavy metals are Cu, Ce and Fe. The single metal that will decrease lung function are Cu and Ce, while Fe has a protective effect on lung function. In view of the serious consequences caused by abnormal lung function, it is crucial to find ways to reduce heavy metals levels in human body.

Environmental implication

The pollution caused by heavy metals to the environment has become a great concern. The concentration of heavy metals in human body could accumulate through a variety of ways. Blood heavy metals levels have been shown to be adversely associated with a series of diseases. However, evidence on heavy metal exposure and lung function is scarce, especially the evidence of joint effects of multiple heavy metal mixtures on lung function. This study provided new evidence regarding the effect of mixed heavy metals exposure on lung function.

Ethics statement

The study was approved by Binzhou Medical University ethics committee (NO.2019075).

CRediT authorship contribution statement

Minghao Wang: Methodology, Investigation, Formal analysis, Writing – original draft, Writing – review & editing. **Lailai Yan:** Methodology, Investigation, Writing – review & editing. **Siqi Dou, Liu Yang:** Investigation. **Yiwen Zhang, Wenzhong Huang:** Writing – review & editing. **Shanshan Li, Peng Lu, Yuming Guo:** Project administration, Methodology, Investigation, Formal analysis, Writing – original draft, Writing – review & editing, Funding acquisition, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

The authors are grateful to all students of Binzhou Medical University who participated in this study. Shanshan Li was supported by an Early Career Fellowship of the Australian National Health and Medical Research Council (APP1109193); Yuming Guo was supported by a career development fellowship of the Australian National Health and Medical Research Council (APP1163693); Peng Lu was sponsored by Taishan Scholar Foundation (tsqn202211228), Shandong Province Natural Science Foundation (ZR202103050697) and Shandong Province Environmental Health Innovative Team.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the


online version at [doi:10.1016/j.jhazmat.2023.132064](https://doi.org/10.1016/j.jhazmat.2023.132064).

References

- [1] Al Osman, M., Yang, F., Massey, I.Y., 2019. Exposure routes and health effects of heavy metals on children. *Biometals* 32, 563–573. <https://doi.org/10.1007/s10534-019-00193-5>.
- [2] Rusyniak, D.E., Arroyo, A., Acciani, J., Froberg, B., Kao, L., Furbee, B., 2010. Heavy metal poisoning: management of intoxication and antidotes. *Exs* 100, 365–396. https://doi.org/10.1007/978-3-7643-8338-1_11.
- [3] Rodríguez Martín, J.A., De Arana, C., Ramos-Miras, J.J., Gil, C., Boluda, R., 2015. Impact of 70 years urban growth associated with heavy metal pollution. *Environ Pollut* 196, 156–163. <https://doi.org/10.1016/j.envpol.2014.10.014>.
- [4] Dunea, D., Iordache, S., Liu, H.Y., Böhler, T., Pohoata, A., Radulescu, C., 2016. Quantifying the impact of PM_{2.5} and associated heavy metals on respiratory health of children near metallurgical facilities. *Environ Sci Pollut Res Int* 23, 15395–15406. <https://doi.org/10.1007/s11356-016-6734-x>.
- [5] Godri Pollitt, K.J., Maikawa, C.L., Wheeler, A.J., Weichenthal, S., Dobbin, N.A., Liu, L., Goldberg, M.S., 2016. Trace metal exposure is associated with increased exhaled nitric oxide in asthmatic children. *Environ Health* 15, 94. <https://doi.org/10.1186/s12940-016-0173-5>.
- [6] Domingo-Relloso, A., Grau-Perez, M., Briongos-Figuero, L., Gomez-Ariza, J.L., Garcia-Barrera, T., Dueñas-Laita, A., Bobb, J.F., Chaves, F.J., Kioumourtzoglou, M. A., Navas-Acien, A., Redon-Mas, J., Martin-Escudero, J.C., Tellez-Plaza, M., 2019. The association of urine metals and metal mixtures with cardiovascular incidence in an adult population from Spain: the Horteaga Follow-Up Study. *Int J Epidemiol* 48, 1839–1849. <https://doi.org/10.1093/ije/dyz061>.
- [7] Lim, J.T., Tan, Y.Q., Valeri, L., Lee, J., Geok, P.P., Chia, S.E., Ong, C.N., Seow, W.J., 2019. Association between serum heavy metals and prostate cancer risk - A multiple metal analysis. *Environ Int* 132, 105109. <https://doi.org/10.1016/j.envint.2019.105109>.
- [8] Reuben, A., Elliott, M.L., Abraham, W.C., Broadbent, J., Houts, R.M., Ireland, D., Knodt, A.R., Poulton, R., Ramrakha, S., Hariri, A.R., Caspi, A., Moffitt, T.E., 2020. Association of childhood lead exposure with MRI measurements of structural brain integrity in midlife. *Jama* 324, 1970–1979. <https://doi.org/10.1001/jama.2020.19998>.
- [9] Shakir, S.K., Aziyullah, A., Murad, W., Daud, M.K., Nabeela, F., Rahman, H., Ur Rehman, S., Häder, D.P., 2017. Toxic metal pollution in pakistan and its possible risks to public health. *Rev Environ Contam Toxicol* 242, 1–60. <https://doi.org/10.1007/978-94-007-6016-9>.
- [10] Duprez, D.A., Jacobs Jr., D.R., 2018. Lung function decline and increased cardiovascular risk: quo vadis. *J Am Coll Cardiol* 72, 1123–1125. <https://doi.org/10.1016/j.jacc.2018.07.015>.
- [11] Madrigal, J.M., Persky, V., Pappalardo, A., Argos, M., 2018. Association of heavy metals with measures of pulmonary function in children and youth: Results from the National Health and Nutrition Examination Survey (NHANES). *Environ Int* 121, 871–878. <https://doi.org/10.1016/j.envint.2018.09.045>.
- [12] Leem, A.Y., Kim, S.K., Chang, J., Kang, Y.A., Kim, Y.S., Park, M.S., Kim, S.Y., Kim, E.Y., Chung, K.S., Jung, J.Y., 2015. Relationship between blood levels of heavy metals and lung function based on the Korean National Health and Nutrition Examination Survey IV-V. *Int J Chron Obstruct Pulmon Dis* 10, 1559–1570. <https://doi.org/10.2147/copd.S86182>.
- [13] Zhou, M., Xiao, L., Yang, S., Wang, B., Shi, T., Tan, A., Wang, X., Mu, G., Chen, W., 2020. Cross-sectional and longitudinal associations between urinary zinc and lung function among urban adults in China. *Thorax* 75, 771–779. <https://doi.org/10.1136/thoraxjnl-2019-213909>.
- [14] Abakay, A., Gokalp, O., Abakay, O., Evliyaoglu, O., Sezgi, C., Palanci, Y., Ekici, F., Karakus, A., Tanrikulu, A.C., Ayhan, M., 2012. Relationships between respiratory function disorders and serum copper levels in copper mineworkers. *Biol Trace Elem Res* 145, 151–157. <https://doi.org/10.1007/s12011-011-9184-9>.
- [15] Wu, L., Cui, F., Ma, J., Huang, Z., Zhang, S., Xiao, Z., Li, J., Ding, X., Niu, P., 2022. Associations of multiple metals with lung function in welders by four statistical models. *Chemosphere* 298, 134202. <https://doi.org/10.1016/j.chemosphere.2022.134202>.
- [16] Sobel, M., Navas-Acien, A., Powers, M., Grau-Perez, M., Goessler, W., Best, L.G., Umans, J., Oelsner, E.C., Podolanczuk, A., Sanchez, T.R., 2022. Environmental-level exposure to metals and metal-mixtures associated with spirometry-defined lung disease in American Indian adults: evidence from the Strong Heart Study. *Environ Res* 207, 112194. <https://doi.org/10.1016/j.envres.2021.112194>.
- [17] Wu, K.G., Chang, C.Y., Yen, C.Y., Lai, C.C., 2019. Associations between environmental heavy metal exposure and childhood asthma: a population-based study. *J Microbiol Immunol Infect* 52, 352–362. <https://doi.org/10.1016/j.jmii.2018.08.001>.
- [18] Qin, G., Niu, Z., Yu, J., Li, Z., Ma, J., Xiang, P., 2021. Soil heavy metal pollution and food safety in China: effects, sources and removing technology. *Chemosphere* 267, 129205. <https://doi.org/10.1016/j.chemosphere.2020.129205>.
- [19] Yu, P., Han, Y., Wang, M., Zhu, Z., Tong, Z., Shao, X., Peng, J., Hamid, Y., Yang, X., Deng, Y., Huang, Y., 2023. Heavy metal content and health risk assessment of atmospheric particles in China: a meta-analysis. *Sci Total Environ* 867, 161556. <https://doi.org/10.1016/j.scitotenv.2023.161556>.
- [20] Miao, J., Feng, S., Wang, M., Jiang, N., Yu, P., Wu, Y., Ye, T., Wen, B., Lu, P., Li, S., Guo, Y., 2022. Life-time summer heat exposure and lung function in young adults: a retrospective cohort study in Shandong China. *Environ Int* 160, 107058. <https://doi.org/10.1016/j.envint.2021.107058>.
- [21] Feng, S., Miao, J., Wang, M., Jiang, N., Dou, S., Yang, L., Ma, Y., Yu, P., Ye, T., Wu, Y., Wen, B., Lu, P., Li, S., Guo, Y., 2022. Long-term improvement of air quality associated with lung function benefits in Chinese young adults: a quasi-experiment cohort study. *Sci Total Environ* 851, 158150. <https://doi.org/10.1016/j.scitotenv.2022.158150>.
- [22] Miller, M.R., Hankinson, J., Brusasco, V., Burgos, F., Casaburi, R., Coates, A., Crapo, R., Enright, P., van der Grinten, C.P., Gustafsson, P., Jensen, R., Johnson, D. C., MacIntyre, N., McKay, R., Navajas, D., Pedersen, O.F., Pellegrino, R., Viegi, G., Wanger, J., 2005. Standardisation of spirometry. *Eur Respir J* 26, 319–338. <https://doi.org/10.1183/09031936.05.00034805>.
- [23] Gosdin, L., Sharma, A.J., Suchdev, P.S., Jefferds, M.E., Young, M.F., Addo, O.Y., 2022. Limits of detection in acute-phase protein biomarkers affect inflammation correction of serum ferritin for quantifying iron status among school-age and preschool-age children and reproductive-age women. *J Nutr* 152, 1370–1377. <https://doi.org/10.1093/jn/nxack035>.
- [24] Cao, B., Yan, L., Ma, J., Jin, M., Park, C., Nozari, Y., Kazmierczak, O.P., Zuckerman, H., Lee, Y., Pan, Z., Brietzke, E., McIntyre, R.S., Lui, L.M.W., Li, N., Wang, J., 2019. Comparison of serum essential trace metals between patients with schizophrenia and healthy controls. *J Trace Elem Med Biol* 51, 79–85. <https://doi.org/10.1016/j.jtemb.2018.10.009>.
- [25] Bobb, J.F., Valeri, L., Claus Henn, B., Christiani, D.C., Wright, R.O., Mazumdar, M., Godleski, J.J., Coull, B.A., 2015. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics* 16, 493–508. <https://doi.org/10.1093/biostatistics/kxu058>.
- [26] Bobb, J.F., Claus Henn, B., Valeri, L., Coull, B.A., 2018. Statistical software for analyzing the health effects of multiple concurrent exposures via Bayesian kernel machine regression. *Environ Health* 17, 67. <https://doi.org/10.1186/s12940-018-0413-y>.
- [27] Wang, Y., Tang, Y., Li, Z., Hua, Q., Wang, L., Song, X., Zou, B., Ding, M., Zhao, J., Tang, C., 2020. Joint toxicity of a multi-heavy metal mixture and chemoprevention in sprague dawley rats. *Int J Environ Res Public Health* 17. <https://doi.org/10.3390/ijerph17041451>.
- [28] Jan, A.K., Moore, J.V., Wang, R.J., McGing, M., Farr, C.K., Moisi, D., Hartman-Filson, M., Kerruish, R., Jeon, D., Lewis, E., Crothers, K., Lederman, M.M., Hunt, P. W., Huang, L., 2021. Markers of inflammation and immune activation are associated with lung function in a multi-center cohort of persons with HIV. *Aids* 35, 1031–1040. <https://doi.org/10.1097/qad.0000000000002846>.
- [29] Kim, S.S., Meeker, J.D., Keil, A.P., Aung, M.T., Bommarito, P.A., Cantonwine, D.E., McElrath, T.F., Ferguson, K.K., 2019. Exposure to 17 trace metals in pregnancy and associations with urinary oxidative stress biomarkers. *Environ Res* 179, 108854. <https://doi.org/10.1016/j.envres.2019.108854>.
- [30] Hancox, R.J., Gray, A.R., Sears, M.R., Poulton, R., 2016. Systemic inflammation and lung function: a longitudinal analysis. *Respir Med* 111, 54–59. <https://doi.org/10.1016/j.rmed.2015.12.007>.
- [31] Okeleji, L.O., Ajayi, A.F., Adebayo-Gege, G., Aremu, V.O., Adebayo, O.I., Adebayo, E.T., 2021. Epidemiologic evidence linking oxidative stress and pulmonary function in healthy populations. *Chronic Dis Transl Med* 7, 88–99. <https://doi.org/10.1016/j.cdtm.2020.11.004>.
- [32] Gehring, U., Beelen, R., Eeftens, M., Hoek, G., de Hoogh, K., de Jongste, J.C., Keuken, M., Koppelman, G.H., Meliefste, K., Oldenwening, M., Postma, D.S., van Rossem, L., Wang, M., Smit, H.A., Brunekreef, B., 2015. Particulate matter composition and respiratory health: the PIAMA Birth Cohort study. *Epidemiology* 26, 300–309. <https://doi.org/10.1097/ede.0000000000000264>.
- [33] Zhang, Z., Weichenthal, S., Kwong, J.C., Burnett, R.T., Hatzopoulou, M., Jerrett, M., van Donkelaar, A., Bai, L., Martin, R.V., Copes, R., Lu, H., Lakey, P., Shiraiwa, M., Chen, H., 2021. A population-based Cohort Study of respiratory disease and long-term exposure to iron and copper in fine particulate air pollution and their combined impact on reactive oxygen species generation in human lungs. *Environ Sci Technol* 55, 3807–3818. <https://doi.org/10.1021/acs.est.0c05931>.
- [34] Chen, Z., Zhou, H., Jia, Y., Ruan, H., Diao, Q., Li, M., Zheng, L., Yao, S., Guo, Y., Zhou, Y., Jiang, Y., 2022. The involvement of copper, circular RNAs, and inflammatory cytokines in chronic respiratory disease. *Chemosphere* 303, 135005. <https://doi.org/10.1016/j.chemosphere.2022.135005>.
- [35] Tanrikulu, A.C., Abakay, A., Evliyaoglu, O., Palanci, Y., 2011. Coenzyme Q10, copper, zinc, and lipid peroxidation levels in serum of patients with chronic obstructive pulmonary disease. *Biol Trace Elem Res* 143, 659–667. <https://doi.org/10.1007/s12011-010-8897-5>.
- [36] Schwoetzer, D., Ernst, H., Schaudien, D., Kock, H., Pohlmann, G., Dasenbrock, C., Creutzenberg, O., 2017. Effects from a 90-day inhalation toxicity study with cerium oxide and barium sulfate nanoparticles in rats. *Part Fibre Toxicol* 14, 23. <https://doi.org/10.1186/s12989-017-0204-6>.
- [37] Rice, K.M., Nalabotu, S.K., Manne, N.D., Kolli, M.B., Nandyala, G., Arvapalli, R., Ma, J.Y., Blough, E.R., 2015. Exposure to cerium oxide nanoparticles is associated with activation of mitogen-activated protein kinases signaling and apoptosis in rat lungs. *J Prev Med Public Health* 48, 132–141. <https://doi.org/10.3961/jpmph.15.006>.
- [38] Ali, M.K., Kim, R.Y., Karim, R., Mayall, J.R., Martin, K.L., Shahandeh, A., Abbasian, F., Starkey, M.R., Loustaud-Ratti, V., Johnstone, D., Milward, E.A., Hansbro, P.M., Horvat, J.C., 2017. Role of iron in the pathogenesis of respiratory disease. *Int J Biochem Cell Biol* 88, 181–195. <https://doi.org/10.1016/j.biocel.2017.05.003>.
- [39] Lee, C.H., Goag, E.K., Lee, S.H., Chung, K.S., Jung, J.Y., Park, M.S., Kim, Y.S., Kim, S.K., Chang, J., Song, J.H., 2016. Association of serum ferritin levels with smoking and lung function in the Korean adult population: analysis of the fourth and fifth Korean National Health and Nutrition Examination Survey. *Int J Chron Obstruct Pulmon Dis* 11, 3001–3006. <https://doi.org/10.2147/copd.S116982>.

- [40] Shaaban, R., Kony, S., Driss, F., Leynaert, B., Soussan, D., Pin, I., Neukirch, F., Zureik, M., 2006. Change in C-reactive protein levels and FEV1 decline: a longitudinal population-based study. *Respir Med* 100, 2112–2120. <https://doi.org/10.1016/j.rmed.2006.03.027>.
- [41] Quezada-Pinedo, H.G., Mensink-Bout, S.M., Reiss, I.K., Jaddoe, V.W.V., Vermeulen, M.J., Duijts, L., 2021. Maternal iron status during early pregnancy and school-age, lung function, asthma, and allergy: the Generation R Study. *Pedia Pulmonol* 56, 1771–1778. <https://doi.org/10.1002/ppul.25324>.
- [42] Lagorio, S., Forastiere, F., Pistelli, R., Iavarone, I., Michelozzi, P., Fano, V., Marconi, A., Ziemacki, G., Ostro, B.D., 2006. Air pollution and lung function among susceptible adult subjects: a panel study. *Environ Health* 5, 11. <https://doi.org/10.1186/1476-069x-5-11>.

Association of psychological distress and DNA methylation: A 5-year longitudinal population-based twin study

Xuanming Hong, MD ^{1,2}, Ke Miao, BS,^{1,2} Weihua Cao, MD,^{1,2} Jun Lv, PhD,^{1,2} Canqing Yu, PhD,^{1,2} Tao Huang, PhD,^{1,2} Dianjianyi Sun, PhD,^{1,2} Chunxiao Liao, PhD,^{1,2} Yuanjie Pang, PhD,^{1,2} Runhua Hu, BS,^{1,2} Zengchang Pang, BS,³ Min Yu, MD,⁴ Hua Wang, MD,⁵ Xianping Wu, MD,⁶ Yu Liu, MD,⁷ Wenjing Gao, PhD^{1,2*} and Liming Li, MD^{1,2*}

Aim: To identify the psychological distress (PD)-associated 5'-cytosine-phosphate-guanine-3' sites (CpGs), and investigate the temporal relationship between dynamic changes in DNA methylation (DNAm) and PD.

Methods: This study included 1084 twins from the Chinese National Twin Register (CNTR). The CNTR conducted epidemiological investigations and blood withdrawal twice in 2013 and 2018. These included twins were used to perform epigenome-wide association studies (EWASs) and to validate the previously reported PD-associated CpGs selected from previous EWASs in PubMed, Embase, and the EWAS catalog. Next, a cross-lagged study was performed to examine the temporality between changes in DNAm and PD in 308 twins who completed both 2013 and 2018 surveys.

Results: The EWAS analysis of our study identified 25 CpGs. In the validation analysis, 741 CpGs from 29 previous EWASs

on PD were selected for validation, and 101 CpGs were validated to be significant at a false discovery rate <0.05. The cross-lagged analysis found a unidirectional path from PD to DNAm at 14 CpGs, while no sites showed significance from DNAm to PD.

Conclusions: This study identified and validated PD-related CpGs in a Chinese twin population, and suggested that PD may be the cause of changes in DNAm over time. The findings provide new insights into the molecular mechanisms underlying PD pathophysiology.

Keywords: DNA methylation, longitudinal studies, psychological distress, twin study.

<http://onlinelibrary.wiley.com/doi/10.1111/pcn.13606/full>

Mental disorders are the leading contributors to the global health-related burden, as they substantially affect daily functioning and quality of life, increase health care costs, and shorten life expectancy.^{1,2} Psychological distress (PD; negative stress) is a general term that refers to nonspecific symptoms of depression, anxiety, and stress.³ As one of the most vital risk factors for developing severe mental disorders, PD causes a significant rise in anxiety and depression. A substantial amount of research has focused on the effects of PD on health, particularly after the outbreak of coronavirus disease 2019 (COVID-19).⁴

The fundamental roles of epigenetic regulation, such as DNA methylation (DNAm), have been implicated in mental disorders such as depression since DNAm can alter depression-related gene expression.⁵ To date, DNAm in several important genes, such as *BDNF*, *NR3C1*, and *FKBP5* genes, are implicated in the regulation of states of depression and anxiety in clinical samples.⁶ As a common feature of both depression and anxiety states, PD is also well documented to be associated with DNAm.⁷ Compelling evidence supports that DNAm may be an essential physiological mechanism for responding to mental stress.⁸ To date, there has been much epigenetic research on PD and related mental disorders, and hundreds of associated DNAm sites (5'-cytosine-phosphate-guanine-3' sites [CpGs]) have been

reported.^{9–11} However, the results of these studies have varied widely, i.e. epigenetic studies on psychology have poor reproducibility.¹² Since various factors, including genetic, psychological, and environmental, can cause mental disorders, controlling for confounding factors presents a significant challenge to these studies in this field.¹³ Twins, particularly monozygotic (MZ) twins, share the same age, genetic information, and some environmental factors. The twin study design is a valuable tool for discovering changes in epigenetic markers associated with diseases.^{14,15}

The plasticity of DNAm may be another explanation for the poor reproducibility.¹⁶ Research suggests that mental stress can cause alterations in DNAm, and through epigenetic regulation, it may have profound effects on psychological and physical disease outcomes or influence the severity of diseases.^{17,18} However, most relevant studies were based on DNAm at specific genes and cross-sectional designs, while longitudinal studies across multiple genes are further needed to validate this mechanism.¹²

In the present investigation, we performed an epigenome-wide association study (EWAS) on PD using whole-blood samples of 1084 twins from the Chinese National Twin Registry (CNTR) and a candidate CpGs association study based on previously reported CpGs associated with PD. Next, we conducted a longitudinal study to investigate

¹ Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing, China

² Key Laboratory of Epidemiology of Major Diseases, Ministry of Education, Peking University, Beijing, China

³ Qingdao Center for Disease Control and Prevention, Qingdao, China

⁴ Zhejiang Center for Disease Control and Prevention, Hangzhou, China

⁵ Jiangsu Center for Disease Control and Prevention, Nanjing, China

⁶ Sichuan Center for Disease Control and Prevention, Chengdu, China

⁷ Heilongjiang Center for Disease Control and Prevention, Harbin, China

* Correspondence: Email: pkuepigwj@126.com; lmlee@vip.163.com

the temporality between DNAm and the levels of PD by adopting a powerful cross-lagged analysis method.

Methods

Study population

The current study was based on data from CNTR. Detailed descriptions of the study design, data collection procedures, and population characteristics have been provided elsewhere.¹⁹ In short, participants from the CNTR were recruited from 11 provinces/cities in China with large-scale baseline and follow-up investigations. The information collection included questionnaires (levels of PD, demographic information, lifestyles, and medical history), blood withdrawal (levels of DNAm and serum biochemical tests), and physical examinations (weight and height). The surveys were conducted twice, in 2013 and 2018. Study participants all provided informed consent. The study protocols were reviewed and approved by the biomedical ethics committee at Peking University, Beijing, China (reference number: IRB00001052-13022, IRB00001052-14021). Our study conforms to the provisions of the Declaration of Helsinki.

Participants who fulfilled the following inclusion criteria were included in the current study: (1) both twins completed questionnaires and physical examinations; (2) blood samples were donated from both twins; and (3) twins completed at least one data collection in 2013 or 2018. Pregnant women and their cotwins were excluded, and if one twin was removed in the following DNAm data quality control or statistical analysis process, the cotwin was then excluded. A total of 1088 participants (mean age, 49.9 years; range, 19 to 82 years) were initially included in the study. Among these, longitudinal data were available for 318 participants (29.2% of the total study population) who completed both 2013 and 2018 investigations.

Measurement of PD, twin zygosity, and covariates

The levels of PD were measured with the 6-Item Kessler Psychological Distress Scale (K6). Participants were enquired about the frequency of feeling nervous, so sad that nothing could cheer them up, restless or fidgety, hopeless, and worthless. The frequency was graded as 0 (all), 1 (most), 2 (some), 3 (a little), and 4 (none of the time). The area under the curve (AUC) of K6 was assessed as 0.86, which was higher than the 0.76 AUC of the World Health Organization's Composite International Diagnostic Interview Short-Form scales.²⁰

An Illumina Methylation Chip panel of 59 single-nucleotide polymorphisms (SNPs) was used to estimate twin zygosity. MZ twins were defined as twins with >80% identical SNPs.²¹ In the present study, 758 twins were identified as MZ twins.

Body mass index (BMI) was calculated as weight in kilograms divided by height in meters squared (kg/m^2). Smoking was categorized as current, former, and never smokers.²² Drinking was classified as current, former, and never drinkers.²³

DNAm measurements

Genomic DNA was extracted from peripheral blood using a BioTeke DNA extraction kit and then bisulfite-converted with an EZ DNA Methylation Kit (Zymo Research). Whole-genome DNAm levels were measured on Illumina 450K or EPIC Human Methylation Arrays, interrogating DNAm at 485,512 and 853,307 CpGs across the genome, respectively (Illumina). Assay reproducibility was 98% between the two BeadChips for samples, and 90% of probes in 450K could be replicated with the EPIC BeadChip.²⁴ Illumina EPIC and 450K samples were merged into a combined data set using the 'combineArrays' function in the R package 'minfi' version 1.34.0,²⁵ and only probes that appeared on both 450K and EPIC microarrays were kept for analysis.

Methylation levels were quantified as β values ranging from 0 to 1, which indicate the proportion of methylation at each CpGs. They were calculated using the formula: $\beta = M/(M + U + 100)$, in which M and U represent the mean probe signal intensity at each site for the methylated and unmethylated states, respectively.²⁵ The probe signal

intensities at each CpGs were measured using R package 'minfi'. After the calculation, β -values were quantile normalized utilizing the R package 'minfi' and adjusted for blood cell counts using 'ChAMP' package version 2.18.3.²⁶ To minimize the impact of potential confounders generated during the DNAm detection procedure, we conducted experimental batch correction using the surrogate variable (SV) analysis function in the R package 'sva' version 3.38.0.²⁷ A total of 121 and 28 SVs were generated for PD in the EWAS and candidate CpGs validation analysis, respectively. The ComBat approach was adopted for longitudinal data since the sample size was relatively small. Quality control and probe exclusions were conducted for DNAm data, and the details of the process are provided in Supplementary Text and Fig. S1.

In total, our study included 378,654 CpGs that appeared on both 450K and EPIC arrays and 1084 participants with DNAm information (308 participants with longitudinal DNAm data) for analysis.

Statistical analysis

An overview of the analysis procedures is provided in Fig. 1. Statistical analyses were conducted using R software version 4.0.3.

Identification of CpGs for PD

We conducted an EWAS and a candidate CpGs association study to identify possible CpGs.

In the EWAS analysis, we assessed the association between PD and methylation levels epigenome-wide for each CpGs included in this study. Candidate CpGs were selected from the previously published literature and were validated for their associations with PD. The online databases PubMed (<https://www.ncbi.nlm.nih.gov/pubmed>), Embase (<https://www.embase.com>), and EWAS catalog (<http://www.ewascatalog.org/>) were systematically searched for previous EWASs on PD and related mental disorders up to June 10, 2023. The detailed search strategy, including selection standards, is described in the Supplementary Materials.

In the full analysis population ($n = 1,084$), we performed linear mixed-effect (LME) models to assess the associations between the DNAm of the CpGs and the levels of PD for both the EWAS and candidate CpGs association study. The DNAm levels at each CpGs were included in the models as the dependent variable, while the K6 score was the continuous independent variable. Age, sex, smoking, drinking, BMI, and SVs (generated from SV analysis) were fixed effects in the LME model. Twin ID and zygosity (MZ or dizygotic [DZ]), shared between the twin pairs, were considered random intercept terms.

A discordant twin analysis was performed in MZ twins to further control the genetic background behind the associations (i.e. within-pair analysis, $n = 758$). We also utilized LME models in this part of the analysis to examine the associations between the K6 score difference and the DNAm difference within MZ twin pairs. The difference was calculated as the trait value of one twin minus that of their cotwin. The average K6 score within twin pairs, in addition to age, sex, smoking, drinking, and BMI, were fixed effects, and twin ID was a random effect term in this analysis. The ComBat method was used to adjust the batch effects of DNAm data for discordant MZ twin analyses.

Sensitivity analysis was performed for both EWAS and candidate CpGs association study by removing those patients who reported antidepressant medication use and their cotwins. All P values were corrected by the false discovery rate (FDR) with the Benjamin-Hochberg procedure. Significant identification and validation were reported for CpGs with an FDR-adjusted $P < 0.05$. To annotate the CpGs, the annotation file for the Illumina Infinium EPIC array was used.

Cross-lagged model analysis

Cross-lagged model analysis was conducted on the longitudinal data to evaluate the temporality between dynamic changes in DNAm and

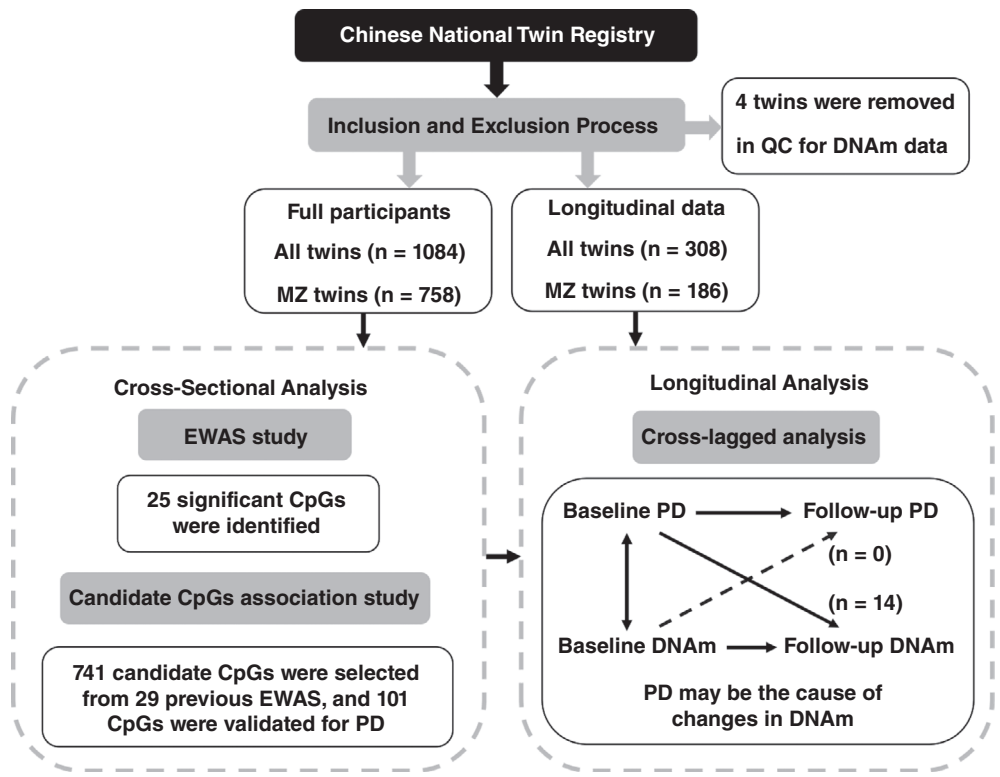


Fig. 1 Analysis workflow. CpGs, 5'-cytosine-phosphate-guanine-3' site; DNAm, DNA methylation; EWAS, epigenome-wide association study; MZ, monozygotic; PD, psychological distress; QC, quality control process.

the levels of PD. The details and the parameter settings of the cross-lagged model are provided in Fig. S2.

The CpGs validated to be associated with PD in the association study were analyzed in the cross-lagged model. This model assesses reciprocal and temporal relationships between traits while accounting for the temporal stability of each construct through time and the concurrent associations between them. Before fitting the models, β -values for DNAm were adjusted for batch effects using the ComBat approach, and then log transformations were performed to improve the fit of the cross-lagged models. In the models, ' ρ_1 ' indicates the prediction from the baseline DNAm level at a specific CpGs to the within-person changes in the levels of PD at follow-up. In contrast, ' ρ_2 ' denotes the prediction of the opposite cross-lagged orientation. Covariates were baseline age, sex, smoking, drinking, and BMI. The comparative fit index (CFI) and standard root mean square residual (SRMR) determined adequate model fit: models with a CFI >0.90 and an SRMR <0.08 were defined as adequately fitted.^{28,29}

The cross-lagged models were established using the R package 'lavaan' version 0.6.15 in twins with longitudinal data ($n = 308$), and the correlations between twin pairs were controlled by the 'cluster' option in the R package.³⁰ Cross-lagged coefficients with $P < 0.05$ were considered significant.

In addition, the results derived from the cross-lagged models were validated by utilizing the Inference About Causation Through Examination of Familial Confounding (ICE FALCON) approach, which allowed us to test the potential causality between DNAm and the levels of PD. The details of the ICE FALCON approach are described elsewhere.³¹ In short, the approach inferred the causality between two phenotypes by assessing the confounding sources for phenotypes of twin pairs using generalized estimation equations (GEEs). Further details of the modeling in this study are provided in the Supplementary Materials (Supplementary Text).

Results

Characteristics of the Study Population

Table 1 describes the detailed demographic information of the participants. The 1084 participants had a mean age of 50.0 years (standard

deviation [SD], 12.2 years) and included 758 MZ twins. The mean K6 score was 7.1 ± 3.3 . The mean difference in K6 score between twins was 1.7 for the full study twins and 1.6 for MZ twins (SDs of

Table 1. Characteristics of the analytic samples by study group

| | Cross-sectional analysis | Longitudinal analysis | |
|------------------------------------|--------------------------|-----------------------|-----------------|
| | Total | Baseline | Follow-up |
| Number | 1084 | 308 | |
| Age, years | 50.0 ± 12.1 | 50.2 ± 10.2 | 54.9 ± 10.2 |
| Women, n (%) | 341 (31.5) | 121 (39.3) | |
| MZ, n (%) | 758 (69.9) | 186 (60.4) | |
| K6 score | 7.1 ± 3.2 | 5.9 ± 2.6 | 8.0 ± 2.8 |
| hs-CRP, mg/L | 1.8 ± 3.8 | 1.6 ± 3.0 | 1.8 ± 3.5 |
| BMI, kg/m ² | 24.8 ± 3.9 | 24.2 ± 3.6 | 24.3 ± 3.5 |
| Smoking status, n (%) | | | |
| Current smoker | 355 (32.7) | 95 (30.8) | 87 (28.2) |
| Former smoker | 143 (13.2) | 25 (8.1) | 35 (11.4) |
| Nonsmoker | 586 (54.1) | 188 (61.0) | 186 (60.4) |
| Alcohol consumption, n (%) | | | |
| Current drinker | 461 (42.5) | 156 (50.6) | 85 (27.6) |
| Former drinker | 73 (6.8) | 10 (3.2) | 27 (8.8) |
| Nondrinker | 550 (50.7) | 142 (46.1) | 196 (63.6) |
| Depression, n (%) [†] | 13 (1.2) | 2 (0.6) | 3 (1.0) |
| Antidepressant medication, n (%) | 7 (0.6) | 1 (0.3) | 1 (0.3) |

[†]Depression was identified by self-report and documented hospital discharge diagnoses.

^a BMI, body mass index; hs-CRP, high-sensitivity C-reactive protein; K6, 6-Item Kessler Psychological Distress Scale; MZ, monozygotic.

Table 2. Cross-lagged coefficients for DNAm and psychological stress in the longitudinal analysis

| Probe ID | CpG _{base} → psychological stress _{follow} | | Psychological stress _{base} → CpG _{follow} | | Goodness of fit | |
|------------|--|----------|--|----------|-----------------|------|
| | $\rho 1$ | <i>P</i> | $\rho 2$ | <i>P</i> | SRMR | CFI |
| cg16360861 | -3.67 | 0.38 | <0.01 | 0.01 | 0.03 | 0.93 |
| cg24222435 | 0.65 | 0.41 | 0.01 | 0.02 | 0.01 | 0.96 |
| cg06117184 | -0.14 | 0.51 | 0.04 | 0.01 | 0.02 | 0.94 |
| cg00080118 | 0.08 | 0.82 | 0.02 | 0.02 | 0.02 | 0.94 |
| cg13423282 | 0.29 | 0.35 | 0.01 | 0.04 | 0.01 | 0.95 |
| cg19159162 | 0.55 | 0.15 | -0.03 | 0.02 | 0.03 | 0.90 |
| cg04275707 | 0.12 | 0.70 | 0.01 | 0.02 | 0.03 | 0.91 |
| cg26224466 | 0.11 | 0.80 | 0.03 | 0.04 | 0.01 | 0.98 |
| cg00102615 | 7.74 | 0.28 | >-0.01 | 0.04 | 0.06 | 0.82 |
| cg22512377 | 0.21 | 0.62 | 0.02 | 0.01 | 0.04 | 0.86 |
| cg10675453 | -0.50 | 0.39 | -0.01 | <0.01 | 0.01 | 0.98 |

$\rho 1$ represents the cross-lagged paths from baseline DNA methylation (DNAm) levels to follow-up psychological stress, and $\rho 2$ indicates the path from baseline psychological stress to follow-up DNAm. CFI, comparative fit index; CpG, 5'-cytosine-phosphate-guanine-3; SRMR, standard root mean square residual.

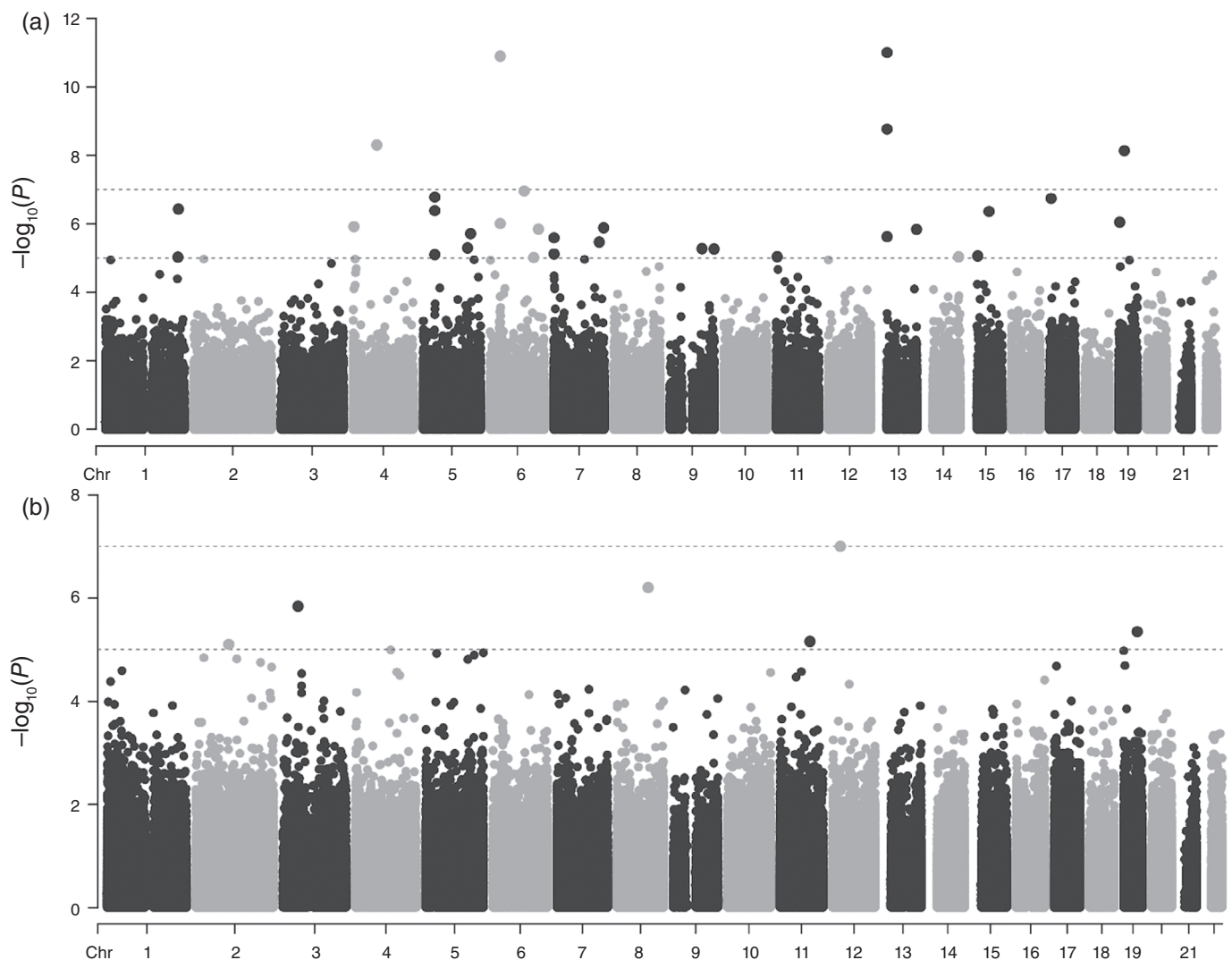


Fig. 2 Manhattan plots of the epigenome-wide association study. Manhattan plots for (a) psychological distress in the full study population and (b) psychological distress in monozygotic twins. The red horizontal lines represent the level of epigenome-wide significance ($P = 1 \times 10^{-5}$ and 1×10^{-7}).

2.6 and 2.5, respectively), and the within-pair correlation was 0.24 in MZ and 0.26 in DZ twins. A total of 13 participants (1.2%) were identified as having depression by self-report or documented hospital discharge diagnoses, and seven participants (0.6%) were taking antidepressant medication.

Identification of CpGs associated with PD

In the EWAS for PD, 24 PD-associated CpGs remained significant after FDR correction for multiple comparisons, while one CpG was identified at FDR <0.05 in discordant MZ twin analysis (cg03625010) (Table S1). The Manhattan plot and QQ plot for EWAS analysis are shown in Figs 2 and S3, respectively. In the sensitivity analysis of the EWAS, a subset of seven patients who were using antidepressant medication were excluded from the analysis with their cotwins (14 patients in total) and 24 sites were identified, which were consistent with the results of the primary analyses in the overall population (Table S2). Conversely, no site reached significance in the sensitivity analysis conducted on discordant MZ twins, where 12 patients were excluded.

Next, to validate the associations between PD and the previously reported CpGs, we searched PubMed, Embase, and EWAS catalog databases for relevant EWAS studies. Figure 3 displays the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2009 flow chart for the literature search process of finding candidate CpGs, which followed PRISMA guidelines.³² After inclusion and exclusion, 29 articles providing 902 candidate CpGs were identified. Among these, 741 sites were available in our DNAm data (Table S3).

For the full analysis population, we validated 85 candidate CpGs to be associated with the levels of PD. Among these, only 11 sites have reported the association direction in the original literature, and all of them revealed the same directions as in previous studies. In the discordant MZ twin analysis, 12 sites were validated for PD, and four sites with the reported association direction by the original literature

all showed the same directions (Table S4). The sensitivity analysis conducted on the candidate CpGs association study yielded validation for 47 sites in the full population analysis and five sites in the analysis of discordant MZ twins. Of these, an additional five and one sites were validated in the sensitivity analysis for the full population and discordant MZ twins, respectively. All of these sites exhibited the same direction of association as reported in the original literature (Table S5).

Finally, a total of 125 CpGs were identified or validated for PD and retained for the subsequent analysis. These CpGs were annotated to 119 genes. Among these CpGs, 94 were at the enhancer or the putative promoter region of the genes (transcriptional start site [TSS] 1500, TSS200, 5' untranslated region, and the first EXON), suggesting a regulatory effect on the expression of the specific genes.^{33,34}

Cross-lagged analysis

The cross-lagged analysis, based on longitudinal twin data ($n = 308$), investigated the temporality between PD and DNAm.

Eleven significant cross-lagged paths were observed for the full analysis population, and all were from the levels of baseline PD to DNAm at follow-up. Among these, two models (from base PD to DNAm of cg00102615 and cg22512377 at follow-up) showed poor fit based on values of CFI (0.82 and 0.86, respectively) <0.90 (Table 2).

We conducted a cross-lagged analysis in MZ twins as a sensitivity analysis. The results were similar to the main results: unidirectional effects from the levels of PD to DNAm were found at 12 CpGs. However, the model fits for five of these sites were poor (cg01947751, cg16360861, cg19825186, cg18163441, and cg16302458) and were removed from the following analysis. In addition, cg22033189, cg03584288, cg10041390, cg15587947, and cg19921130 were identified in MZ twins (Table S6). In summary, we identified significant

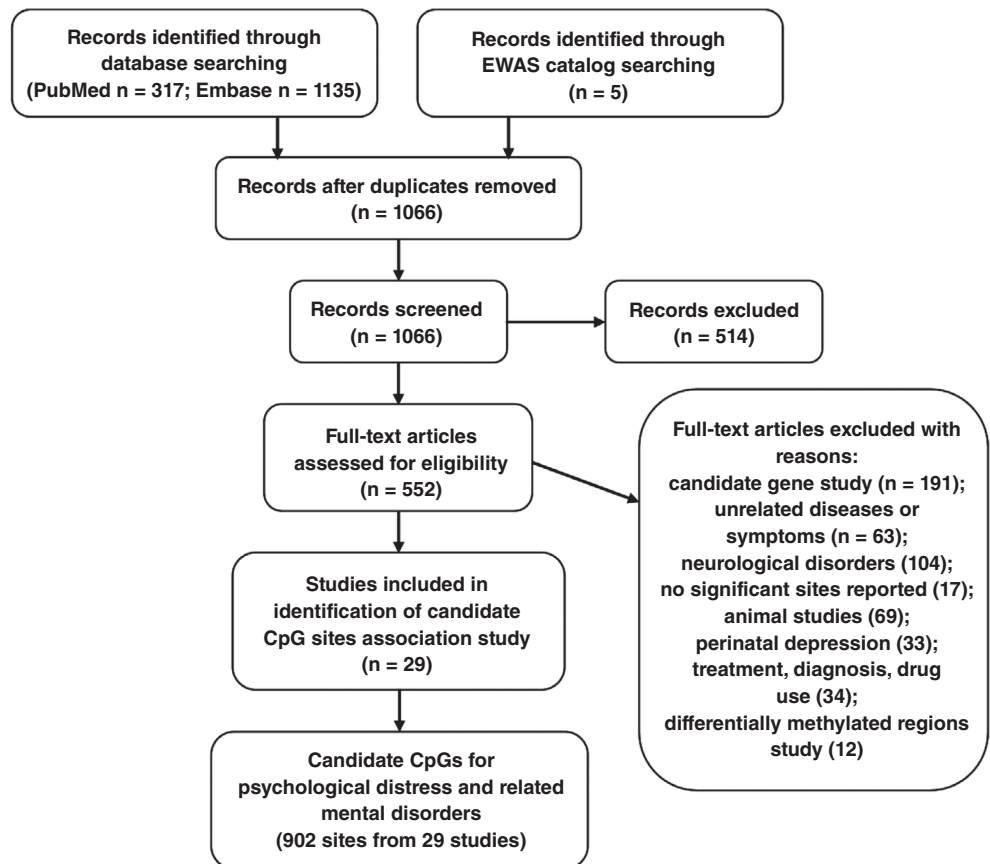


Fig. 3 Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2009 flow chart. CpGs, 5-cytosine-phosphate-guanine-3' site; EWAS, epigenome-wide association study.

cross-lagged effects from baseline PD to follow-up DNAm at 14 CpGs. No CpGs showed a path from DNAm levels at baseline to the levels of PD at follow-up.

Finally, ICE FALCON analysis was conducted to validate whether the temporal relationships from the cross-lagged analysis could be causal. The causal role of PD on DNAm at cg24222435 was validated (Table S7). The model settings and the interpretation of the results are presented in detail in the Supplementary Materials (Supplementary Text).

Discussion

In the current study, we first conducted an EWAS on PD and identified 25 associated CpGs. Then, in the candidate CpGs association study, we systematically searched the PubMed, Embase, and EWAS catalog databases and validated 101 CpGs for PD and related mental disorders. The longitudinal analysis revealed 14 cross-lagged predictions, all from the baseline levels of PD to DNAm at follow-up. Among these, the causal effect of PD on the methylation of cg24222435 was validated using the ICE FALCON approach.

CpGs associated with PD

In the EWAS analysis, we identified 24 novel CpGs for PD, which were annotated to 14 genes. Among these, four CpGs were annotated at the enhancer or the putative promoter region of genes, and these genes have been implicated in mental health. cg13720581 was annotated to the tenascin-XB gene (*TNXB*), which have been linked genetically and epigenetically with psychiatric disorders, such as schizophrenia and anorexia nervosa in previous studies.^{35,36} The *TRPV5* gene (annotated from cg13675849) may participate in the pathophysiology of several cerebral diseases, such as schizophrenia and depression, through the role of serum- and glucocorticoid-inducible kinase 1 in the regulation of neuronal function.³⁷ The *NRG2* gene (annotated from cg11061655) encodes a novel member of the neuregulin family of growth and differentiation factors.³⁸ An animal study used *NRG2* knockout mice to assess the function of this gene and revealed a novel role for *NRG2* in the modulation of behaviors relevant to psychiatric disorders.³⁹ cg03625010, identified in the discordant MZ analysis, was annotated to the *SOX5* gene, which participate in neurogenesis and other discrete developmental processes by encoding one of the SOX family of transcription factors that are involved in cell fate and differentiation.⁴⁰ Notably, three sites from the EWAS results were annotated to the *N6AMT2* gene, which is known to be involved in the methyl transfer process and is over-expressed in many cancers, which may reflect the increased protein synthesis needs of fast-growing cells.⁴¹

Although there have been many studies on the associations between DNAm and PD and related mental disorders, and hundreds of DNAm sites or regions have been reported, the results from different studies are inconsistent. A recent systematic review noted that the reproducibility of epigenetic studies on psychology was poor, especially for individual CpGs.¹² Various factors could contribute to the inconsistency of results, including small sample size, lack of standardized objective assessments of PD and statistical analysis. The inadequate adjustment for key covariates such as blood cell counts and batch effects and the less stringent threshold for detection ($P < 1e-4$ or $FDR < 0.10$ rather than $FDR < 0.05$) could both affect the outcomes and lead to low credibility of the results.^{42,43} In addition, the methylation level and the psychological status can both be influenced by ethnic differences and might thus have resulted in confounding that could bias estimates.⁴⁴⁻⁴⁶ To validate these associations that have been previously reported and provide evidence from the Chinese population, an association study of candidate CpGs in Chinese twins was conducted after controlling for genetic and environmental factors, as well as batch effects and blood cell composition.

The 90 CpGs validated in the candidate CpGs association study for the full analysis population were annotated to 99 genes. Among these, 12 CpGs were at the genes' enhancer or putative promoter region and have been implicated in psychological disorders.

cg19014730, cg00610228, and cg00052684 were all annotated to the *FKBP5* gene. This gene encodes the protein FKBP5 (FK506-binding protein 5), which regulates the sensitivity of the glucocorticoid receptor and is critically involved in the response mechanisms to stress and anxiety disorders.^{47,48} Studies have indicated that DNAm of the *FKBP5* gene is assumed to alter *FKBP5* expression and hence the synthesis of FKBP5.⁴⁹ Our study further provides evidence that methylation of the *FKBP5* gene may play a significant role in physiological and pathological processes of PD. A previous review reported that behavioral abnormalities, including aggression and anxiety, are frequently observed in patients carrying pathogenic variants of *HNRNPU* gene (annotated from cg22320000).⁵⁰ *GCH1* (annotated from cg05105845) is suggested to be associated with anxiety and pain-related processes.⁵¹ Variations in *NTRK2* (annotated from cg13965062) are highlighted as potential depression risk factors.⁵² Downregulation of *HSPA1B* (annotated from cg19159162) may alter glucocorticoid sensitivity and has been conceived as an offender in the pathophysiology of major depressive disorder (MDD).⁵³ Moreover, previous studies have reported that depression-like behaviors were observed in mice with overexpression of the *PTEN* gene (annotated from cg10041390), and the deletion or knockdown of *PTEN* in the prefrontal cortex prevented depression-like behaviors.⁵⁴ The expression of the *PGM1* gene (annotated from cg02994863) was reported to be increased in patients with bipolar disorder.⁵⁵ The *MBOAT7* gene (annotated from cg12173535) is primarily expressed in the liver and is related to inflammatory processes.⁵⁶ Pathogenic variants of this gene are newly discovered to be a rare cause of intellectual disability and autism spectrum disorder.⁵⁷ The *CNNM2* gene (annotated from cg23843362) has been reported to have a pivotal role in brain development.⁵⁸ A variant rs7914558 at *CNNM2* was associated with brain structure and cognition, suggesting an association between *CNNM2* and PD.⁵⁹ Decreased *NDUFA* (annotated from cg07987587) expression in lymphocytes of patients with bipolar disorder has been reported by previous studies evaluating markers of mitochondrial oxidative phosphorylation.⁶⁰ The upregulation of the *VPS35* gene (annotated from cg09238957) has been shown to rescue α -synuclein-induced neurodegeneration and has been linked to Parkinson disease and depressive-like behaviors.^{61,62} The *GGA3* gene (annotated from cg09238957) is highly expressed in the brain and neurons. Previous animal experiments have confirmed the role of *GGA3* in novelty-induced hyperactivity and decreased anxiety-like behaviors.⁶³ An experimental animal study demonstrated that the gene expression of *CITED2* (annotated from cg05607246 in the sensitivity analysis) can be altered by chronic psychological stress.² Another study examining the molecular basis of mood disorders reported that *CITED2* is one of the hub genes, while its expression and dysregulation patterns were associated with mood disorder.⁶⁴

In the discordant MZ analysis, genetic background and early environmental factors were further controlled, and 13 CpGs were validated. It is worth mentioning that 11 of the 13 sites were additionally validated in the MZ twin population but not in the full analysis population. One reason for this fact is that since discordant MZ twin design controls for genetic variation, environmental factors are suggested to play a major role in these validated CpGs. A well-documented effect of the environment on DNAm has been demonstrated to contribute to disease susceptibility through methylation.⁸ The findings from discordant MZ analysis may provide new insight into the environmental contributions underlying the relationship between PD and DNAm.

From the results, the *NEGR1* gene (annotated from cg09256413), which is highly expressed in the cerebral cortex and hippocampus, was one of the most significant genes for depression.⁶⁵ Brain-derived neurotrophic factor (*BDNF*, annotated from cg23497217), as one of the major neurotrophic factors, plays an essential role in the survival, differentiation, and growth of peripheral and central neurons during development and in adulthood.⁶⁶ Of note, many studies have demonstrated the relationships of *BDNF* with psychological disorders, including depression, distress, and anxiety.⁶⁷ In addition to the above-discussed

genes, many genes, such as *RNF8* annotated from cg27597069, *DNAJA2* from cg06939115, and *SNHG12* from cg22033189, were reported to be associated with neurodevelopmental processes, nerve signal transmission, and psychiatric and neurodegenerative diseases, which may underlie common pathophysiology in various psychological disorders.^{68–74}

In addition, 68 of the 101 CpGs validated from the candidate CpGs association study overlapped with an EWAS on MDD based on 39 Japanese individuals (including 20 patients with MDD and 19 controls) from Tokushima and Kochi University Hospitals, with a mean age of 44.2 ± 15.2 years.⁷⁵ The similar ethnic backgrounds and age range of our study population may be the leading cause for the overlap, suggesting that ethnic and genetic factors and age play a part in the epigenetic mechanisms of PD.

Temporality between DNAm and PD

The effects of epigenetic modulation on the development of psychological disorders have been the subject of intense study. Increasing evidence indicates that physiological and psychological stress may alter DNAm at critical genes.⁵ The Environmental Risk (E-Risk) Longitudinal Twin Study in the United Kingdom has made several contributions in the field of DNAm change resulting from psychological stress, childhood adverse experiences and childhood psychotic symptoms.^{12,76,77}

To our knowledge, this is the first study that used a cross-lagged design to investigate the temporal relationship between DNAm and PD. Consistent with previous studies, we revealed that the levels of PD could predict DNAm at 14 CpGs, suggesting the causal effects of the levels of PD on epigenetic alterations. By adopting the ICE FALCON approach, we validated the causal relationship between PD and the methylation level at cg24222435. PD is the cause of several lasting biological consequences, particularly for the endocrine system. It can also affect a range of intermediate phenotypes, including brain structure and function, physiological function, and behavioral changes.⁷⁸ Evidence suggests that molecular mechanisms related to epigenetic regulation may contribute to susceptibility and resilience to the effects of trauma and stress.⁸

On the other hand, through certain pathways, DNAm may increase the impact of stressful experiences. A study based on longitudinal data from 100 middle-aged Black women used structural equation models and found that DNAm at *OXTR* may mediate the associations between adult adversity and the development of depression.⁷⁹ Another study on 33 adolescents aged 12 to 13 years observed that DNAm might mediate the relationships between neighborhood disadvantage and brain development.⁸⁰ Although further validation is needed, this evidence demonstrates that the DNAm influenced by distress might play a role in developing more severe mental/neurological diseases.

Our study provides strong evidence that PD could act on the alteration of DNAm. However, the specific roles and mechanisms underlying the effects remain poorly understood, and further exploration is needed.

Strengths and limitations

The present study has strengths. First, we adopted two strategies to identify the PD-related CpGs: EWAS analysis and a candidate CpGs association study were conducted among the twin population, which added confidence to the observed results and helped increase the probability of identifying significant CpGs for the subsequent analysis. For DNAm studies, the twin population has a unique value because they are naturally matched for genetic and environmental factors. Second, to our knowledge, this is the first study to assess the temporality between DNAm and the levels of PD by studying their cross-lagged effect patterns.

The study also has several limitations. First, the K6 scale score is not a diagnostic measure, and it includes only depression and anxiety symptoms, while the dimensions of severe psychopathologies are

not demonstrated. However, the K6 score may be a better indicator of the need for mental health services than individual disorders that vary widely in severity since it is a dimensional scale.²⁰ Second, the associations between PD and DNAm at a number of CpGs previously reported were not validated in the present study, and may be due to ethnic heterogeneity of methylation levels and their relationship with PD. Different study designs may also be a potential explanation. Further validation is needed for other ethnicities and ancestries. Finally, the sample size for longitudinal and discordant MZ twin analysis was relatively small. A larger sample size is required to investigate the association and temporality between DNAm and the levels of PD.

Conclusion

In summary, our study identified 25 CpGs in the EWAS analysis and validated 101 CpGs from previous studies in the Chinese twin population for their associations with PD. Using the longitudinal design, we found a unidirectional effect of the levels of PD on DNAm at 14 CpGs, and a causal effect of PD on DNAm at cg24222435 was validated, suggesting that PD is likely to be the cause of changes in DNAm over time. These findings provide new insights into the molecular mechanisms underlying the pathophysiology of PD.

Acknowledgments

The authors gratefully acknowledge the support of the National Natural Science Foundation of China (82073633, 81973126, 81573223) and the Special Fund for Health Scientific Research in Public Welfare (201502006, 201002007), and Peking University Outstanding Discipline Construction Project of Epidemiology and Biostatistics.

Disclosure statement

The authors declare that they have no competing interests.

Author contributions

X.H. conducted the statistical analysis and was the main contributor in writing the manuscript. K.M. checked the contents of this manuscript. W.C., J.L., C.Y., R.H., W.G., and L.L. performed the data collection and reviewed the manuscript. T.H., D.S., C.L., and Y.P. reviewed the contents of this manuscript. Z.P., M.Y., H.W., X.W., and Y.L. recruited the participants for the study.

Ethics statement

This study was approved by the biomedical ethics committee at Peking University, Beijing, China (IRB00001052-13022, IRB00001052-14021, IRB00001052-22032). Informed consent was obtained from all recruited participants.

References

1. Santomauro DF, Mantilla Herrera AM, Shadid J *et al.* Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. *Lancet*. 2021; **398**: 1700–1712.
2. Zyla J, Kabacik S, O'Brien G *et al.* Combining CDKN1A gene expression and genome-wide SNPs in a twin cohort to gain insight into the heritability of individual radiosensitivity. *Funct. Integr. Genomics* 2019; **19**: 575–585.
3. Javadi Arjmand E, Bemanian M, Vold JH *et al.* Emotional eating and changes in high-sugar food and drink consumption linked to psychological distress and worries: A cohort study from Norway. *Nutrients* 2023; **15**: 778.
4. Holmes EA, O'Connor RC, Perry VH *et al.* Multidisciplinary research priorities for the COVID-19 pandemic: A call for action for mental health science. *Lancet Psychiatry* 2020; **7**: 547–560.
5. Yao B, Cheng Y, Wang Z *et al.* DNA N6-methyladenine is dynamically regulated in the mouse brain following environmental stress. *Nat Commun* 2017; **8**: 1122.
6. Bagot RC, Labonte B, Pena CJ *et al.* Epigenetic signaling in psychiatric disorders: Stress and depression. *Dialogues Clin. Neurosci.* 2014; **16**: 281–295.
7. Dahl C, Guldberg P. DNA methylation analysis techniques. *Biogerontology* 2003; **4**: 233–250.








8. Fraga MF, Ballestar E, Paz MF *et al.* Epigenetic differences arise during the lifetime of monozygotic twins. *Proc. Natl. Acad. Sci. U. S. A.* 2005; **102**: 10604–10609.
9. Story Jovanova O, Nedeljkovic I, Spieler D *et al.* DNA methylation signatures of depressive symptoms in middle-aged and elderly persons: Meta-analysis of multiethnic epigenome-wide studies. *JAMA Psychiatry* 2018; **75**: 949–959.
10. Shen X, Caramaschi D, Adams MJ *et al.* DNA methylome-wide association study of genetic risk for depression implicates antigen processing and immune responses. *Genome Med.* 2022; **14**: 36.
11. Howard DM, Pain O, Arathimos R *et al.* Methylome-wide association study of early life stressors and adult mental health. *Hum. Mol. Genet.* 2022; **31**: 651–664.
12. Zhang Y, Liu C. Evaluating the challenges and reproducibility of studies investigating DNA methylation signatures of psychological stress. *Epigenomics* 2022; **14**: 405–421.
13. Cheng Y, Sun M, Chen L *et al.* Ten-eleven translocation proteins modulate the response to environmental stress in mice. *Cell Rep* 2018; **25**: 3194–3203.
14. Wang Z, Peng H, Gao W *et al.* Blood DNA methylation markers associated with type 2 diabetes, fasting glucose, and HbA1c levels: An epigenome-wide association study in 316 adult twin pairs. *Genomics* 2021; **113**: 4206–4213.
15. Dempster EL, Pidsley R, Schalkwyk LC *et al.* Disease-associated epigenetic changes in monozygotic twins discordant for schizophrenia and bipolar disorder. *Hum. Mol. Genet.* 2011; **20**: 4786–4796.
16. Martin DI, Cropley JE, Suter CM. Epigenetics in disease: Leader or follower? *Epigenetics* 2011; **6**: 843–848.
17. Hannum G, Guinney J, Zhao L *et al.* Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell* 2013; **49**: 359–367.
18. LaPlant Q, Vialou V, Covington HE 3rd *et al.* Dnmt3a regulates emotional behavior and spine plasticity in the nucleus accumbens. *Nat. Neurosci.* 2010; **13**: 1137–1143.
19. Gao W, Cao W, Lv J *et al.* The Chinese National Twin Registry: A ‘gold mine’ for scientific research. *J. Intern. Med.* 2019; **286**: 299–308.
20. Olsson M, Wang S, Wall M, Marcus SC, Blanco C. Trends in serious psychological distress and outpatient mental health care of US adults. *JAMA Psychiatry* 2019; **76**: 152–161.
21. Wang B, Gao W, Yu C *et al.* Determination of zygosity in adult Chinese twins using the 450K methylation Array versus questionnaire data. *PLoS ONE* 2015; **10**: e0123992.
22. Mathur R, Rentsch CT, Morton CE *et al.* Ethnic differences in SARS-CoV-2 infection and COVID-19-related hospitalisation, intensive care unit admission, and death in 17 million adults in England: An observational cohort study using the OpenSAFELY platform. *Lancet.* 2021; **397**: 1711–1724.
23. Effeo VS, Correa A, Chen H, Lacy ME, Bertoni AG. High-sensitivity C-reactive protein is associated with incident type 2 diabetes among African Americans: The Jackson heart study. *Diabetes Care* 2015; **38**: 1694–1700.
24. Pidsley R, Zotenko E, Peters TJ *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol* 2016; **17**: 208.
25. Aryee MJ, Jaffe AE, Corrada-Bravo H *et al.* Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 2014; **30**: 1363–1369.
26. Tian Y, Morris TJ, Webster AP *et al.* ChAMP: Updated methylation analysis pipeline for Illumina BeadChips. *Bioinformatics* 2017; **33**: 3982–3984.
27. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 2012; **28**: 882–883.
28. Mayberry LS, Osborn CY. Empirical validation of the information-motivation-behavioral skills model of diabetes medication adherence: A framework for intervention. *Diabetes Care* 2014; **37**: 1246–1253.
29. Deschenes SS, Graham E, Kivimaki M *et al.* Adverse childhood experiences and the risk of diabetes: Examining the roles of depressive symptoms and cardiometabolic dysregulations in the Whitehall II cohort study. *Diabetes Care* 2018; **41**: 2120–2126.
30. Rosseel Y. Lavaan: an R Package for structural equation modeling. *J. Stat. Softw.* 2012; **48**: 1–36.
31. Li S, Bui M, Hopper JL. Inference about causation from examination of familial confounding (ICE FALCON): A model for assessing causation analogous to mendelian randomization. *Int. J. Epidemiol.* 2020; **49**: 1259–1269.
32. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med.* 2009; **6**: e1000097.
33. Gutierrez-Arcelus M, Lappalainen T, Montgomery SB *et al.* Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife* 2013; **2**: e00523.
34. Li K, Qin L, Jiang S *et al.* The signature of HBV-related liver disease in peripheral blood mononuclear cell DNA methylation. *Clin. Epigenetics* 2020; **12**: 81.
35. Wei J, Hemmings GP. TNXB locus may be a candidate gene predisposing to schizophrenia. *Am J Med Genet B Neuropsychiatr Genet* 2004; **125b**: 43–49.
36. Booij L, Casey KF, Antunes JM *et al.* DNA methylation in individuals with anorexia nervosa and in matched normal-eater controls: A genome-wide study. *Int. J. Eat. Disord.* 2015; **48**: 874–882.
37. Lang F, Strutz-Seeböhm N, Seeböhm G, Lang UE. Significance of SGK1 in the regulation of neuronal function. *J. Physiol.* 2010; **588**: 3349–3354.
38. Wilson KJ, Mill CP, Cameron EM, Hobbs SS, Hammer RP, Riese DJ II. Inter-conversion of neuregulin2 full and partial agonists for ErbB4. *Biochem. Biophys. Res. Commun.* 2007; **364**: 351–357.
39. Yan L, Shamir A, Skirzewski M *et al.* Neuregulin-2 ablation results in dopamine dysregulation and severe behavioral phenotypes relevant to psychiatric disorders. *Mol. Psychiatry* 2018; **23**: 1233–1243.
40. Zhang F, Rao S, Baranova A. Shared genetic liability between major depressive disorder and osteoarthritis. *Bone Joint Res.* 2022; **11**: 12–22.
41. Hamey JJ, Winter DL, Yagoub D, Overall CM, Hart-Smith G, Wilkins MR. Novel N-terminal and lysine methyltransferases that target translation elongation factor 1A in yeast and human. *Mol. Cell. Proteomics* 2016; **15**: 164–176.
42. Dias S, Adam S, Rheeder P, Louw J, Pfeiffer C. Altered genome-wide DNA methylation in peripheral blood of south African women with gestational diabetes mellitus. *Int. J. Mol. Sci.* 2019; **20**: 5828.
43. Wu HC, Wang Q, Yang HI, Tsai WY, Chen CJ, Santella RM. Global DNA methylation in a population with aflatoxin B1 exposure. *Epigenetics* 2013; **8**: 962–969.
44. Zhang FF, Morabia A, Carroll J *et al.* Dietary patterns are associated with levels of global genomic DNA methylation in a cancer-free population. *J. Nutr.* 2011; **141**: 1165–1171.
45. Zhang M, He Y, Zhang X, Zhang M, Kong L. A pooled analysis of the diagnostic efficacy of plasmic methylated septin-9 as a novel biomarker for colorectal cancer. *Biomed Rep.* 2017; **7**: 353–360.
46. Richardson RA, Keyes KM, Medina JT, Calvo E. Sociodemographic inequalities in depression among older adults: Cross-sectional evidence from 18 countries. *Lancet Psychiatry* 2020; **7**: 673–681.
47. Sawamura T, Klengel T, Armario A *et al.* Dexamethasone treatment leads to enhanced fear extinction and dynamic Fkbp5 regulation in amygdala. *Neuropsychopharmacology* 2016; **41**: 832–846.
48. Binder EB, Bradley RG, Liu W *et al.* Association of FKBP5 polymorphisms and childhood abuse with risk of posttraumatic stress disorder symptoms in adults. *Jama* 2008; **299**: 1291–1305.
49. Klinger-König J, Hertel J, Van der Auwera S *et al.* Methylation of the FKBP5 gene in association with FKBP5 genotypes, childhood maltreatment and depression. *Neuropsychopharmacology* 2019; **44**: 930–938.
50. Taylor J, Spiller M, Ranguin K *et al.* Expanding the phenotype of HNRNPU-related neurodevelopmental disorder with emphasis on seizure phenotype and review of literature. *Am. J. Med. Genet. A* 2022; **188**: 1497–1514.
51. George SZ, Wu SS, Wallace MR *et al.* Biopsychosocial influence on shoulder Pain: Influence of genetic and psychological combinations on twelve-month postoperative Pain and disability outcomes. *Arthritis Care Res (Hoboken)*. 2016; **68**: 1671–1680.
52. Li D, Li Y, Chen Y *et al.* Neuroprotection of reduced thyroid hormone with increased estrogen and progesterone in postpartum depression. *Biosci. Rep.* 2019; **39**: BSR20182382.
53. Schiweck C, Claes S, Van Oudenhove L *et al.* Childhood trauma, suicide risk and inflammatory phenotypes of depression: Insights from monocyte gene expression. *Transl. Psychiatry* 2020; **10**: 296.
54. Wang XQ, Zhang L, Xia ZY, Chen JY, Fang Y, Ding YQ. PTEN in prefrontal cortex is essential in regulating depression-like behaviors in mice. *Transl. Psychiatry* 2021; **11**: 185.
55. Beckman G, Beckman L, Cedergren B, Perris C, Strandman E. Serum protein and red cell enzyme polymorphisms in affective disorders. *Hum. Hered.* 1978; **28**: 41–47.
56. Sookoian S, Flichman D, Garaycochea ME *et al.* Lack of evidence supporting a role of TMC4-rs641738 missense variant-MBOAT7- intergenic

- downstream variant-in the susceptibility to nonalcoholic fatty liver disease. *Sci. Rep.* 2018; **8**: 5097.
57. Jacher JE, Roy N, Ghaziuddin M, Innis JW. Expanding the phenotypic spectrum of MBOAT7-related intellectual disability. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 2019; **180**: 483–487.
58. Thyme SB, Pieper LM, Li EH *et al.* Phenotypic landscape of schizophrenia-associated genes defines candidates and their shared functions. *Cell* 2019; **177**: 478–491.e20.
59. Rose EJ, Hargreaves A, Morris D *et al.* Effects of a novel schizophrenia risk variant rs7914558 at CNNM2 on brain structure and attributional style. *Br. J. Psychiatry* 2014; **204**: 115–121.
60. Machado-Vieira R, Zanetti MV, Teixeira AL *et al.* Decreased AKT1/mTOR pathway mRNA expression in short-term bipolar disorder. *Eur. Neuropsychopharmacol.* 2015; **25**: 468–473.
61. Hussein A, Guevara CA, Del Valle P *et al.* Non-motor symptoms of Parkinson's disease: The neurobiology of early psychiatric and cognitive dysfunction. *Neuroscientist* 2023; **29**: 97–116.
62. Appel JR, Ye S, Tang F *et al.* Increased microglial activity, impaired adult hippocampal neurogenesis, and depressive-like behavior in microglial VPS35-depleted mice. *J. Neurosci.* 2018; **38**: 5949–5968.
63. Walker KR, Modgil A, Albrecht D *et al.* Genetic deletion of the Clathrin adaptor GGA3 reduces anxiety and alters GABAergic transmission. *PLoS One* 2016; **11**: e0155799.
64. Yang C, Zhang K, Zhang A, Sun N, Liu Z, Zhang K. Co-expression network modeling identifies specific inflammation and neurological disease-related genes mRNA modules in mood disorder. *Front. Genet.* 2022; **13**: 865015.
65. Howard DM, Adams MJ, Shirali M *et al.* Genome-wide association study of depression phenotypes in UK biobank identifies variants in excitatory synaptic pathways. *Nat. Commun.* 2018; **9**: 1470.
66. Zhang JC, Yao W, Hashimoto K. Brain-derived neurotrophic factor (BDNF)-TrkB signaling in inflammation-related depression and potential therapeutic targets. *Curr. Neuropharmacol.* 2016; **14**: 721–731.
67. Guo SL, Tan GH, Li S *et al.* Serum inducible kinase is a positive regulator of cortical dendrite development and is required for BDNF-promoted dendritic arborization. *Cell Res.* 2012; **22**: 387–398.
68. Valnegri P, Huang J, Yamada T *et al.* RNF8/UBC13 ubiquitin signaling suppresses synapse formation in the mammalian brain. *Nat Commun* 2017; **8**: 1271.
69. Xie MJ, Ishikawa Y, Yagi H *et al.* PIP3-Phldb2 is crucial for LTP regulating synaptic NMDA and AMPA receptor density and PSD95 turnover. *Sci. Rep.* 2019; **9**: 4305.
70. Mizutani R, Yamauchi J, Kusakawa S *et al.* Sorting nexin 3, a protein upregulated by lithium, contains a novel phosphatidylinositol-binding sequence and mediates neurite outgrowth in N1E-115 cells. *Cell. Signal.* 2009; **21**: 1586–1594.
71. Rosales-Hernandez A, Beck KE, Zhao X, Braun AP, Braun JEA. RDJ2 (DNAJA2) chaperones neural G protein signaling pathways. *Cell Stress Chaperones* 2009; **14**: 71–82.
72. Groffen AJ, Friedrich R, Brian EC *et al.* DOC2A and DOC2B are sensors for neuronal activity with unique calcium-dependent and kinetic properties. *J. Neurochem.* 2006; **97**: 818–833.
73. Coulter ME, Musaev D, DeGennaro EM *et al.* Regulation of human cerebral cortical development by EXOC7 and EXOC8, components of the exocyst complex, and roles in neural progenitor cell proliferation and survival. *Genet. Med.* 2020; **22**: 1040–1050.
74. Zhao M, Wang J, Xi X, Tan N, Zhang L. SNHG12 promotes angiogenesis following ischemic stroke via regulating miR-150/VEGF pathway. *Neuroscience* 2018; **390**: 231–240.
75. Numata S, Ishii K, Tajima A *et al.* Blood diagnostic biomarkers for major depressive disorder using multiplex DNA methylation profiles: Discovery and validation. *Epigenetics* 2015; **10**: 135–141.
76. Ouellet-Morin I, Wong CC, Danese A *et al.* Increased serotonin transporter gene (SERT) DNA methylation is associated with bullying victimization and blunted cortisol response to stress in childhood: A longitudinal study of discordant monozygotic twins. *Psychol. Med.* 2013; **43**: 1813–1823.
77. Fisher HL, Murphy TM, Arseneault L *et al.* Methylomic analysis of monozygotic twins discordant for childhood psychotic symptoms. *Epigenetics* 2015; **10**: 1014–1023.
78. Wang Y, Jiang P, Tang S *et al.* Left superior temporal sulcus morphometry mediates the impact of anxiety and depressive symptoms on sleep quality in healthy adults. *Soc. Cogn. Affect. Neurosci.* 2021; **16**: 492–501.
79. Simons RL, Lei MK, Beach SRH, Cutrona CE, Philibert RA. Methylation of the oxytocin receptor gene mediates the effect of adversity on negative schemas and depression. *Dev. Psychopathol.* 2017; **29**: 725–736.
80. Wrigglesworth J, Ryan J, Vijayakumar N, Whittle S. Brain-derived neurotrophic factor DNA methylation mediates the association between neighborhood disadvantage and adolescent brain structure. *Psychiatry Res Neuroimaging.* 2019; **285**: 51–57.

Supporting Information

Additional supporting information can be found online in the Supporting Information section at the end of this article.

A high-resolution haplotype-resolved Reference panel constructed from the China Kadoorie Biobank Study

Canqing Yu ^{1,2,3,†}, Xianmei Lan ^{4,5,†}, Ye Tao ^{5,†}, Yu Guo ^{6,†}, Dianjianyi Sun ^{1,2,3,†}, Puyi Qian ^{7,†},
Yuwen Zhou ^{4,5,†}, Robin G. Walters ^{8,9}, Linxuan Li ^{4,5}, Yunqing Zhu ¹, Jingyu Zeng ^{5,10},
Iona Y. Millwood ^{8,9}, Ruidong Guo ⁵, Pei Pei ², Tao Yang ⁷, Huaidong Du ^{8,9}, Fan Yang ⁷, Ling Yang ^{8,9},
Fangyi Ren ⁷, Yiping Chen ^{8,9}, Fengzhen Chen ⁷, Xiaosen Jiang ^{4,5}, Zhiqiang Ye ⁷, Lanlan Dai ⁷,
Xiaofeng Wei ⁷, Xun Xu ^{5,11}, Huanming Yang ^{5,12,13}, Jian Wang ¹⁴, Zhengming Chen ^{8,9},
Huanhuan Zhu ^{5,*}, Jun Lv ^{2,15,*}, Xin Jin ^{5,16,*} and Liming Li ^{1,2,3,*}

¹Department of Epidemiology and Biostatistics, School of Public Health, Peking University Health Science Center, Beijing 100191, China

²Center for Public Health and Epidemic Preparedness and Response, Peking University, Beijing 100191, China

³Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing 100191, China

⁴College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

⁵BGI Research, Shenzhen 518083, China

⁶National Center for Cardiovascular Diseases, Fuwai Hospital, Chinese Academy of Medical Sciences, Beijing 100037, China

⁷China National GeneBank, BGI, Shenzhen 518083, China

⁸Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU), Nuffield Department of Population Health, University of Oxford, Oxford OX3 7LF, United Kingdom

⁹Medical Research Council Population Health Research Unit, Nuffield Department of Population Health, University of Oxford, Oxford OX3 7LF, United Kingdom

¹⁰College of Life Sciences, Northwest A&F University, Yangling, Shaanxi 712100, China

¹¹Guangdong Provincial Key Laboratory of Genome Read and Write, BGI Research, Shenzhen 518083, China

¹²Guangdong Provincial Academician Workstation of BGI Synthetic Genomics, BGI, Shenzhen 518083, China

¹³James D. Watson Institute of Genome Sciences, Hangzhou 310013, China

¹⁴BGI, Shenzhen 518083, China

¹⁵State Key Laboratory of Vascular Homeostasis and Remodeling, Peking University, Beijing 100191, China

¹⁶School of Medicine, South China University of Technology, Guangzhou 510006, China

*To whom correspondence should be addressed. Liming Li. Tel: +86 10 82801528; Email: lmlee@vip.163.com

Correspondence may also be addressed to Xin Jin. Tel: +86 15814045013; Email: jinxin@genomics.cn

Correspondence may also be addressed to Jun Lv. Tel: +86 10 82801528; Email: epi.lvjun@vip.163.com

Correspondence may also be addressed to Huanhuan Zhu. Tel: +86 13714302012; Email: zhuhuanhuan1@genomics.cn

†These authors contributed equally.

Abstract

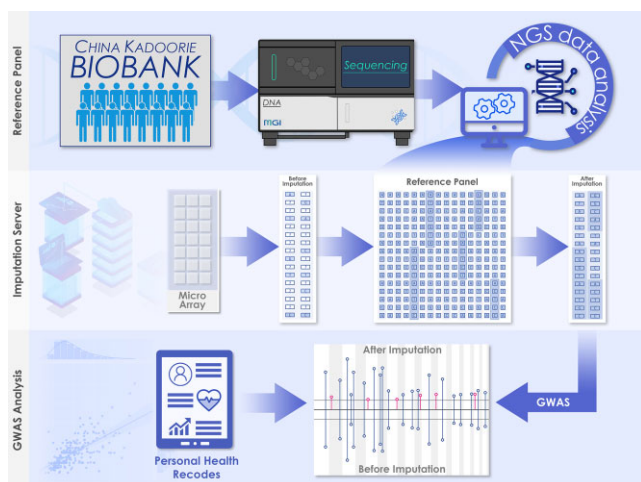
Precision medicine depends on high-accuracy individual-level genotype data. However, the whole-genome sequencing (WGS) is still not suitable for gigantic studies due to budget constraints. It is particularly important to construct highly accurate haplotype reference panel for genotype imputation. In this study, we used 10 000 samples with medium-depth WGS to construct a reference panel that we named the CKB reference panel. By imputing microarray datasets, it showed that the CKB panel outperformed compared panels in terms of both the number of well-imputed variants and imputation accuracy. In addition, we have completed the imputation of 100 706 microarrays with the CKB panel, and the after-imputed data is the hitherto largest whole genome data of the Chinese population. Furthermore, in the GWAS analysis of real phenotype height, the number of tested SNPs tripled and the number of significant SNPs doubled after imputation. Finally, we developed an online server for offering free genotype imputation service based on the CKB reference panel (<https://db.cngb.org/imputation/>). We believe that the CKB panel is of great value for imputing microarray or low-coverage genotype data of Chinese population, and potentially mixed populations. The imputation-completed 100 706 microarray data are enormous and precious resources of population genetic studies for complex traits and diseases.

Received: January 19, 2023. Revised: August 2, 2023. Editorial Decision: August 30, 2023. Accepted: September 12, 2023

© The Author(s) 2023. Published by Oxford University Press on behalf of Nucleic Acids Research.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Graphical abstract



Introduction

In recent years, precision medicine has made remarkable achievements in complex diseases retreatment and development of target drugs by using molecular biological information (e.g. individual genome) and clinical symptoms (1,2). Precision medicine relies on high-throughput whole-genome data to implement individual-based clinical diagnosis and treatment for patients. However, although the cost of whole-genome sequencing (WGS) technology has been greatly reduced, there is still a budget problem for large-scale population research. Most researchers still prefer the low-cost microarray-based genotyping technology, which sequences known loci to obtain genotype data for follow-up analysis. But microarray cannot mine novel mutation sites related to the disease, so there are limitations in the interpretation of the genetic mechanism of the disease. At present, the common method is to impute the microarray data at the whole-genome level based on the appropriate reference panel thus to obtain the whole-genome data for a population. The selection of reference genome plays an important role in the imputation accuracy of genome data and subsequent analysis results.

Internationally, the haplotype map (HapMap) (3,4), 1000 Genomes Project (1KGP) (5,6), the Haplotype reference consortium (HRC) (7) and trans-omics for precision medicine (TOPMed) (8) have been launched. The HapMap project is the next major human genomic program after the International Human Genome Project. In 2007, the HapMap (phase 3) sequenced 1184 individuals from 11 populations. In 2015, American, British and Chinese scientists jointly announced the completion of the Thousand Human Genome Project (phase 3), which sequenced the whole genomes of 2504 individuals from 26 global populations and created the most comprehensive genetic polymorphism map of the human genome. The 1KGP panel is the most-commonly used genome data to date. Recently, the expanded 1KGP cohort including 602 trios were published, in which all 3202 samples were sequenced to a high depth of 30 times (6). In 2016, the HRC project integrated 20 studies, such as UK10K and 1KGP, and created a reference panel with 32 470 individuals mostly with low-coverage WGS data (7). The latest TOPMed reference panel collected 97 256 individuals, including 47 159 Europeans, 24 267 Africans, 17

085 admixed Americans, 1184 East Asians, 644 South Asians and other populations (8).

In recent years, in addition to the international haplotype reference projects, national haploid genome sequence consortiums have also been initiated in various countries, including Netherlands, Denmark, Iceland and Singapore. The Dutch Human Genome Project sequenced 250 pedigrees at moderate depth (12 \times) to construct haploid reference sequences, substantially improving the accuracy of genotype inference for low-frequency variants (9). The Danish Genome Project sequenced 50 Danish families at high-depth (80 \times) WGS to construct the first Danish genome-wide high-precision haplotype reference panel (10). The Icelandic Genome performed high-depth (20 \times) WGS on \sim 2000 individuals to create haplotype reference sequences, significantly improving the efficacy of association analysis and complex disease studies (11). The SG10K reference panel sequenced 4810 individuals, including 2780 Chinese, 903 Malays and 1127 Indians, with an average sequencing depth of 13.7 \times (12). This database is a valuable resource to advance the genetic study of complex traits and diseases in Asians.

China has the largest population in the world, producing enormous genetic resources, and should make a greater contribution to human genetics and complex disease research. However, the lack of high-quality haplotype reference sequences has become a bottleneck in the fields of population genetics and molecular biology. Fortunately, in the past 2 years, researchers have constructed reference panels based on Chinese population: the ChinaMAP (China Metabolic Analytics Project) and the Nyuwa reference panels. The ChinaMAP consortium performed 40 \times deep WGS on 10588 individuals collected from different regions and ethnicities in China (13,14). The library construction and WGS were performed on the BGISEQ-500 platform at BGI-Genomics. The ChinaMAP reference panel is a high-quality genetic variation database of Chinese population and plays an essential role in the analysis of Chinese population structure, genetic variation spectrum and pathogenic variants. The NyuWa reference panel includes 2902 independent samples with high-depth (26.2 \times) WGS collected from 23 administrative regions of China (15). It is important to expand the diversity of genetic resources and

improve the accuracy of medical research in Chinese population.

The China Kadoorie Biobank (CKB), previously known as the Kadoorie Study of Chronic Disease in China (KSCDC), is an international collaborative research project on chronic diseases jointly conducted by Peking University, Chinese Academy of Medical Sciences and University of Oxford, UK (16). It is a gargantuan prospective study and the largest Chinese population cohort to date. During 2004–2008, >510 000 adults were recruited from 10 geographically defined regions in China. The study aims to establish a database of blood samples and clinical information and to investigate the main genetic and environmental causes of common chronic diseases. To date, the CKB cohort has achieved numerous influential findings in clinical studies, such as the relationship between smoking, physical activity, fresh fruit intake, egg consumption and the risk of cardiovascular disease (17–20), the association between diabetes and the risk of death (21) and the relationship between smoking, alcohol and tea consumption and esophageal cancer (22). However, unfortunately, there are no large-scale population genetics and genetic background studies of complex traits and diseases based on the CKB cohort (23). A major reason is the lack of high-density genetic data. Although microarray testing (Affymetrix Axiom myDesign) of >100 000 samples has been completed, the data are still not comparable to WGS data in terms of the number of genetic variants and the detection of novel loci.

In this work, we constructed a high-resolution haplotype-resolved reference panel based on 9950 individuals from the CKB cohort and 50 Chinese samples from the 1KGP study, with an average sequencing depth of 15.41 \times . We evaluated the imputation performance of the CKB reference panel from the perspective of number of imputed variants and imputation accuracy. The compared reference panels include the extended high coverage 1KGP, the newly developed TOPMed, the ChinaMAP and the NyuWa panels built from the Chinese population. In addition, based on the constructed CKB panel, we completed the genotype imputation for 100 706 microarray samples and obtained the largest whole genome data in the Chinese population. We further performed the genome-wide association study (GWAS) of human height based on the 100 706 microarray data before and after imputation. The total number of SNPs used in GWAS tripled after imputation and the number of significant loci increased from 119 to 147, while 26 out of the additional 28 identified loci were previously reported to be associated with height. We also created an online imputation server to offer free genotype imputation service (<https://db.cngb.org/imputation/>).

Materials and methods

Subjects

In this project, we constructed a haplotype reference panel based on 10 000 Chinese individuals, including 9950 from the CKB cohort and 50 from the 1KGP Han Chinese. The CKB project recruited >510 000 adults aged from 30 to 79 in 10 (five urban, five rural) geographic regions of China. These 9950 individuals were stroke cases from the cohort. The 50 1KGP samples included 20 northern and 30 southern Han Chinese. We also used 100 706 CKB microarray samples (independent of the 9950 samples) in subsequent analyses. Writ-

ten informed consent was obtained from all participants from the CKB cohort.

DNA samples and library construction

The WGS was performed for the 10 000 samples. Specifically, DNA concentration was measured by ExKubit dsDNA HS Assay Kits (Shanghai ExCell Biology, Inc) and Fluostar Omega Microplate Reader (BMG Labtech GmbH). The DNA quality was evaluated by agarose gel electrophoresis at a constant voltage (180 V) for 35 min. The DNA shearing was done by the Covaris E220 ultrasonics DNA shearing instruments. The DNA purification and fragment size selection were applied by VAHTS DNA Clean Beads (Vazyme, #N411). The libraries were constructed on BGI's DNBseq-T1 \times 4RS platform and the loading DNA concentration was >12 ng/ μ l. The paired-end 100-bp (PE100) WGS with 350-bp insert sizes was performed on the MGI DNBSEQ sequencing platform.

Variant calling and sample quality control

To perform variant calling on each sample (also known as individual variant calling), we first applied SOAPnuke (v.2.1.1; -n 0.1 -l 12 -M 2) (24) to filter low quality reads and remove adapter sequences. Then, we obtained aligned Binary Alignment/Map (BAM) files by aligning sequence reads to the GRCh38 human reference genome assembly with Sentieon (v.202010.04) bwa-mem algorithm (<https://www.biorxiv.org/content/10.1101/115717v2>). On the sorted and aligned BAM files, we used Sentieon drivers LocusCollector to collect information on duplicates and Dedup to remove the duplicates. For regions that contain insertions or deletions (INDELs), we further performed local realignment around INDELs to correct for mapping errors and increase the quality of INDEL detection by using the Sentieon Realigner algorithm. To increase the accuracy of variant calling, we carried out base quality score recalibration (BQSR) to BAM files based on the Sentieon QualCal algorithm, which created a recalibration table. This table file was then applied as an input to Sentieon Haplotyper for single-nucleotide polymorphisms (SNPs) and INDELs detection. After all these steps, we obtained the called variant sites for each sample in gVCF format. Note that, for this variant calling workflow, we used the Sentieon DNASEq toolkit instead of the GATK best practice (25) for the following reasons: (1) the DNASEq and GATK have near-identical variant detection accuracy, (2) the DNASEq is >30 times faster than GATK and (3) the DNASEq may be more suitable for less deeply sequenced samples (<https://www.biorxiv.org/content/10.1101/115717v2>).

Before performing joint variant calling, we first selected samples with (1) no evidence of contamination (VerifyBamID FREEMIX <0.03) (26), (2) high library quality measured by reads duplication rate <0.05, (3) mean sequencing depth \geq 10 \times and (4) GC content between 40 and 44. The joint variant calling was then performed by GVCFTyper algorithm implemented in Sentieon, followed by variant quality score recalibration (VQSR) for SNPs and INDELs separately using GATK (27). In this way, we first built the models with VariantRecalibrator and then applied it in ApplyVQSR. After that, ExcessHet >54.69 and low-quality sites that did not pass VQSR were filtered out by SelectVariants. Finally, we calculated genotype posterior probabilities by CalculateGenotypePosteriors.

Reference panel construction

After calculated genotype posterior probabilities, we further set low quality genotypes ($GQ < 20$) as missing and then removed low-complexity sites with minimum count of less than one or with missing alternate (ALT) allele. We also split a multiallelic SNP with more than one ALT allele to biallelic SNPs, with each ALT allele in a separate row. Next, we performed genotype phasing (also known as phasing/haplotype estimation), which is the process of statistical estimation of haplotypes from genotype data. This step was done by Beagle v.5.2 (28). Note that, during these steps, we did not remove related samples since the genetic relatedness can be modeled and improve haplotype phase accuracy. This concept was borrowed from the generation of the latest version of the 1000 Genome Project reference panel, in which the phasing accuracy was evaluated between inclusion and exclusion of trios; and the evaluation result showed that phasing with pedigree data achieved higher accuracy compared to unrelated samples alone (6). Finally, we removed close relatives up to the second degree generated by KING v.2.2.7 (29) as the related samples can distort the population allele frequency estimation in the subsequent analysis. After then, we obtained the reference panel, which we named as CKB reference panel. The construction workflow is provided in Figure 1.

We performed annotation analysis with the Ensembl Variant Effect Predictor (VEP) (30) by using plug-ins SIFT (31) and PolyPhen (32) algorithms. Additionally, we used ClinVar (33), (34) to label pathogenic variants and their related diseases in the ClinVar database. We kept variants only when its reference allele and alternate allele were consistent with that of CLIN-HGVS, which is a new INFO tag that reports the top-level genomic HGVS (Human Genome Variation Society) expression for the variant. We further calculated the alternate allele frequency (AF) as $AF = AC/AN$, where AC is the alternate allele count and AN is the total number of alleles.

Evaluation of the imputation performance

We conducted extensive scenarios to evaluate the imputation performance of the CKB panel and others, including the extended 1KGP, TOPMed, ChinaMAP and NyuWa reference panels. There were two datasets to be imputed: the CKB microarray data and the 1KGP microarray data. From the CKB cohort, 50 randomly selected samples independent of that in the CKB reference panel were genotyped in both SNP array and high coverage WGS (44.14×). The 50 1KGP microarray samples were all Chinese and also independent of those in the CKB reference panel. To evaluate the imputation performance, we compared number of imputed variants and imputation accuracy. For the imputed variants, we defined high-quality variants with an imputed information score > 0.8 and medium-quality variants with an imputed information score between 0.4 and 0.8. For the imputation accuracy, we calculated Pearson correlation coefficient (R^2), precision and sensitivity. The high coverage WGS data were treated as ground truth when computing imputation accuracy between the imputed and true genotypes. For the CKB and extended 1KGP panels, we performed imputation procedures locally; while for TOPMed, ChinaMAP and NyuWa, we submitted jobs to their online imputation servers and downloaded the after-imputed files.

To assess the precision and sensitivity, we first calculated the true positive (TP), false positive (FP), false negative (FN) and

true negative (TN). The TP indicates that the imputed genotype correctly predicts the true WGS genotype. The FP is an error classification where the imputed genotype incorrectly indicates the presence of a WGS variant. The FN is also an error classification where the imputed genotype incorrectly indicates the absence of a WGS variant. The TN is an outcome where the predicted genotype correctly predicts the case of homozygous reference calls. The details of 3×3 confusion matrix of defining TP, FP, FN and TN were provided in Supplementary Table S1. To eliminate the bias caused by the number of imputed variants, we compared the ratios of TP, FP, and FN instead of their counts directly. The TN value for all panels was zero. The ratio of TP was calculated by $TP/(TP + FP + FN + TN)$, same for FP and FN. The precision was computed by $TP/(TP + FP)$ and the sensitivity was computed by $TP/(TP + FN)$.

Imputation for 100 706 microarray data

We imputed 100 706 CKB microarray data based on the CKB reference panel in Beagle v.5.2 (28). Note that, the 50 samples with both microarray and high-coverage WGS data were included in the 100 706 individuals. To carry out imputation efficiently, we randomly divided the 100 706 samples into 21 chunks, in which 20 chunks contained 4800 samples and one chunk contained 4706 samples. Then, we parallelly executed genotype imputation for these chunks. To assess the performance for imputing such a large volume of data using the developed CKB panel, we extracted the 50 after-imputed microarray samples and calculated the Pearson correlation coefficients with their high coverage WGS set. We also compared this imputation accuracy with that of imputing the 50 microarray samples alone.

PCA of the CKB reference panel and 100706 microarray data

To detect population stratification, we carried out principal component analysis (PCA) (35,36) of genotype data in the CKB reference panel. The PCA was carried out in Plink v.1.9 (37) with autosomal biallelic SNPs satisfying the following conditions (1) $MAF \geq 1\%$, (2) genotyping rate $\geq 90\%$, (3) Hardy–Weinberg equilibrium (HWE) P -value $> 1E-06$ and (4) low linkage disequilibrium (LD, $r^2 < 0.5$) with other variants in windows of 50 SNPs with steps of five SNPs. In addition, we performed PCA for 100 706 microarray data before imputation. The Plink arguments were the same as used previously.

GWAS analysis of simulated data

In this section, we aimed to perform GWAS of simulated phenotypic values, whereas the genotype data were a combination of the CKB reference panel and after-imputed 100 706 microarray data. First, we performed PCA of genotype data by using PCAone (<https://github.com/Zilong-Li/PCAone>), which was applicable for large samples. Then, we simulated phenotypic data under null and alternative hypotheses, separately. Under the null hypothesis that none of the SNPs were associated with the phenotype, we generated a vector of phenotypic values from a standard normal distribution. Under the alternative hypothesis that the phenotype data was generated from a linear regression model by using five SNPs as independent variables with randomly assigned effects size β . The causal SNPs included rs3003378 ($\beta = 0.02$), rs6764623

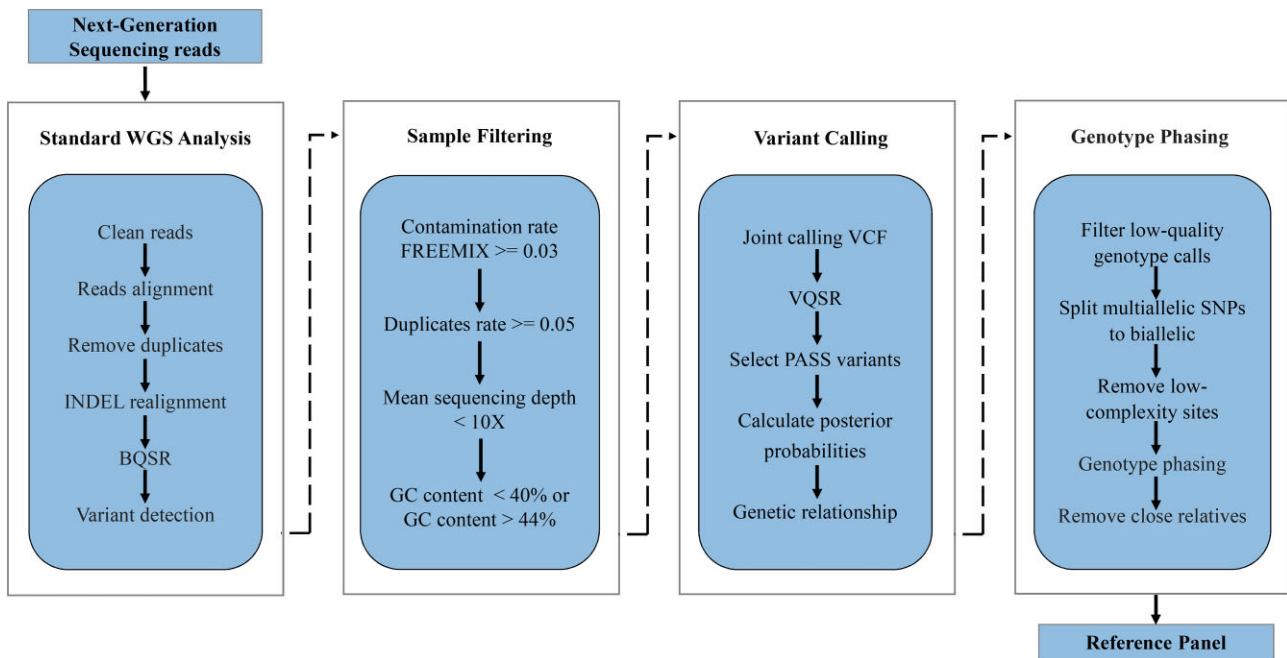


Figure 1. The workflow of panel construction.

($\beta = 0.01$), rs10905649 ($\beta = 0.02$), rs13254191 ($\beta = 0.03$) and rs10915307 ($\beta = 0.01$). We used PC1 to PC5 and sex of the participants as the covariates to carry out GWAS analysis in Plink v.2.0 (38). We reported Manhattan plots, QQ plots, histograms and regional plot for the GWAS results.

GWAS analysis of real phenotype data

In this section, we performed GWAS analysis of real phenotype height, while the genotype data was 100 706 microarray data before and after imputation, separately. The covariates included age, sex, sampling site and the first 10 principal components of the microarray data before imputation. We used Plink v.2.0 for GWAS analysis by testing SNPs with MAF >0.01, HWE P -value >1E-06 and genotype missing rate <0.01. For the GWAS results, we defined a SNP as significant if its P -value >5E-08. We further grouped these significant SNPs into different loci by sliding a fixed-width (1 MB) window. For two loci identified before and after imputation, if the distance between their centers is within 500 KB, we defined that they were a shared locus.

Online imputation service

We developed an online imputation server to offer genotype imputation service, which allows users to run imputation tasks free and safely in an easy way. For the online server, we provided the CKB and 1KGP as available reference panels, GRCh37 (hg19) and GRCh38 (hg38) as human genome assembly, Minimac v.4 (39,40) and Beagle v.5.2 (41) as imputation tools, and different population options. Specifically, for the CKB panel, Chinese is the sole population, and for the 1KGP panel, the available populations include East Asian, South Asian, African, European, Admixed American and all populations. Users can access the server via <https://db.cngb.org/imputation/>.

Results

Data quality

After sample-level quality control, the haplotype reference panel included 9964 individuals, where 9914 were from the CKB cohort and 50 were 1KGP Chinese. The sequencing depth, sex distribution, and age distribution are provided in Figure 2a–c. In detail, the mean sequencing depth was 15.41 (15.41 for CKB samples and 15.78 for 1KGP samples). There were 4416 males (44.32%) and 5548 females (55.68%) in the panel; while specifically in the CKB and 1KGP cohort, the percentages of males were 44.29 and 50.00%, respectively. The sex distribution of the CKB individuals was highly consistent with that in the entire CKB cohort (i.e. male: 41%, female: 59%). We also provide the number of samples received from each sampling site in Figure 2d. Specifically, Heilongjiang, Henan and Guangxi were the top three provinces with the largest recruitments. The other provinces had relatively similar sample sizes. The sex and age distributions of samples in each sampling site are provided in Figure 2e.

We provided a comprehensive comparison in terms of sample size, averaged sequencing depth, number of variants and ancestries between the CKB reference panel and other four panels (Table 1). In detail, the TOPMed is the largest one with sample size 97 256, followed by the ChinaMAP and CKB with ~10 000; the extended 1KGP and NyuWa included ~3000 individuals. The sequencing depth is either medium coverage (10–30 \times) or high coverage (>30 \times). The TOPMed panel has 308.11 million variants, including 286.07 million SNPs and 22.04 million INDELS. The CKB panel had 129.74 million variants, including 113.73 million SNPs and 16.01 INDELS. The ChinaMAP, extended 1KGP, and NyuWa performed variant filtering from database to reference panel. Specifically, the ChinaMAP panel involved SNPs only (59.01 million). The extended 1KGP panel included 70.77 million variants, while SNPs counted 87.21%. The NyuWa panel had 19 million variants. By contrast, the CKB reference panel had relatively

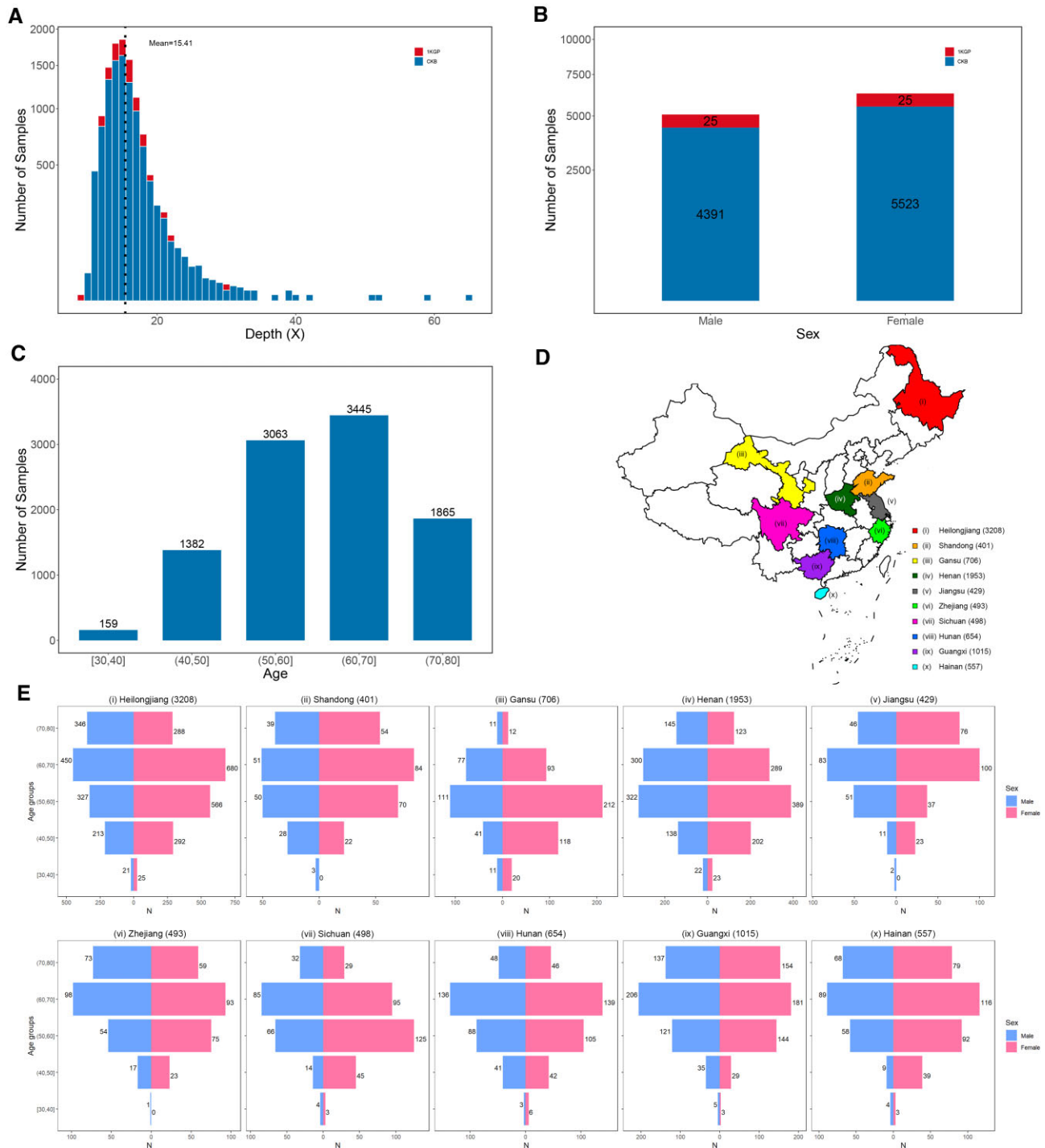


Figure 2. The sample information in the CKB reference panel. **(A)** The sequencing depth distribution of 50 1KGP samples (red) and 9914 CKB samples (blue). The mean sequencing depth of all 9964 samples was 15.41 \times . **(B)** The sex distribution of 50 1KGP samples (red) and 9914 CKB samples (blue). **(C)** The age distribution of 9914 CKB samples. **(D)** The China map colored for 10 sampling sites with number of samples. The total number of samples from all sampling sites was 9914. **(E)** The sex and age distributions of samples in each sampling site.

large sample sizes and detected variants compared with other panels.

In addition, we calculated three quality indicators for SNPs: the heterozygous:homozygous (het:hom) ratio, the transition:transversion (Ti:Tv) ratio and the non-reference genotype concordance rate (NRC). The het:hom ratio is highly dependent on ancestry and the median value for Asians is ~ 1.4 (42). The Ti:Tv ratio reflected the quality of SNP calling and the ex-

pected ratio would be close to 2.0 for human WGS data (27). For the CKB reference panel, we obtained a het:hom ratio of 1.31 and a Ti:Tv ratio of 1.97, indicating the high quality of genotypic data in the constructed panel. The NRC is genotype-aware recall (also known as sensitivity = $TP/(TP + FN)$). We used the genotype data of 50 1KGP samples with high-depth sequencing as actual status and their SNP calls in the CKB panel as the predicted data. The NRC for these

Table 1. The information of CKB and other reference panels

| Reference panel | Sample size | Sequencing depth | Variants | SNP | INDEL | Ancestries |
|-----------------|-------------|------------------|-------------|-------------|------------|---------------------|
| CKB | 9964 | 15.41× | 129 743 542 | 113 731 044 | 16 012 498 | Chinese |
| ChinaMAP | 10 155 | 40.8× | 590 10 860 | 59 010 860 | 0 | Chinese |
| Extended 1KGP | 3202 | 34× | 707 68 225 | 61 715 567 | 9 052 658 | Multiple ancestries |
| TOPMed | 97 256 | >30× | 308 107 085 | 286 068 980 | 22 038 105 | Multiple ancestries |
| NyuWa | 2902 | 26.2× | 19 256 267 | - | - | Chinese |

50 samples were calculated before and after genotype phasing implemented by Beagle v.5.2 (41). The average NRC increased from 0.9811 to 0.9927 and the improvement is more significant for samples with lower sequencing depth (Supplementary Figure S1a).

The PCA of individuals' genotype data in the CKB reference panel is provided in Supplementary Figure S1b. The first PC represents a latitudinal gradient, from north to south China. As expected, individuals in the CKB reference panel were sampled from 10 different regions.

Novel variants and variant annotation

We defined novel variants that were not assigned a unique variant accession identifier (RS number) in dbSNP (Single Nucleotide Polymorphism Database, build 154) (43). Thereby, the number of novel SNPs and INDELs are 50.16 million (44.1%) and 5.42 million (33.8%), respectively (Supplementary Figure S2a and b). Note that, a site with different mutation variety compared to that in dbSNP (e.g. in panel: REF:ALT is A:-, while in dbSNP REF:ALT is CA:C) was also considered as a novel variant, which partially explained the relatively high proportion of novel sites (44,45). As expected, most novel SNPs (99.99%) and INDELs (99.15%) were rare variants (MAF < 0.5%).

Based on the results of VEP annotation analysis, 55% were intronic variants and 26% variants located in the intergenic region. The subsequent categories were non-coding variants (15%), upstream/downstream transcript variants (12%), regulatory variants (4%), variants in mRNA untranslated regions (1%), functional variants (0.8%), transcription factor binding sites (0.3%) and splice-site variants (0.1%) (Supplementary Figure S2c). Among the functional variants, the most abundant class is missense mutation. Based on the ClinVar annotation results, there were 1604, 411, 516, 83, 12 and nine pathogenic variants for AC = 1, AC = 2, AF < 0.1, AF < 1, AF < 5 and AF > 5%, respectively (Supplementary Figure S2d). Specifically, there were nine common pathogenic variants (i.e. alternate allele AF > 5%), including seven single nucleotide variation (SNV), one insertion (INS) and one deletion (DEL) (Table S2). The seven SNVs included rs7417106 (A > G, AF = 0.9468, gnomAD.EAS AF = 0.9429), rs5082 (G > A, AF = 0.9229, gnomAD.EAS AF = 0.9049), rs2280789 (A > G, AF = 0.3531, gnomAD.EAS AF = 0.3285), rs2280788 (G > C, AF = 0.1182, gnomAD.EAS AF = 0.1117), rs3754413 (C > T, AF = 0.0737, gnomAD.EAS AF = 0.0731), rs72474224 (C > T, AF = 0.0522, gnomAD.EAS AF = 0.0854) and rs77592601 (C > T, AF = 0.0510, gnomAD.EAS AF = 0.0479). In correspondence to these SNVs, the ClinVar annotated diseases included renal tubular epithelial cell apoptosis, familial hypercholesterolemia, human immunodeficiency virus type 1, rare genetic deafness, myeloproliferative neoplasm and premature

rupture of membranes. The INS and DEL corresponded to hepatocellular carcinoma. The SIFT and PolyPhen algorithms provided consistent prediction of deleterious variants, that was a large fraction (96%) were very rare variants (AC < 2) (Supplementary Figure S2e). Over 72% variants can be predicted as deleterious by both algorithms (Supplementary Figure S2f). For the low-frequency and common variants (MAF > 0.005), 23 (0.3%) of them were annotated as deleterious. In particular, seven variants were predicted as deleterious mutations by both SIFT and PolyPhen algorithms, seven variants were uniquely annotated by SIFT and nine were uniquely annotated by PolyPhen (Supplementary Table S3).

Imputation performance evaluation

We compared the imputation performance of the CKB reference panel with that of the extended 1KGP (6), TOPMed (46), ChinaMAP (14) and NyuWa (47) from the perspective of number of imputed variants and imputation accuracy. We used 50 CKB and 50 1KGP microarray datasets as input samples to be imputed. The corresponding high-coverage WGS data were used as ground truth datasets. In imputation of the CKB array data, the CKB reference panel provided the highest number of medium-quality imputed variants (10.86 million), followed by the extended 1KGP (10.01 million), NyuWa (9.23 million), TOPMed (8.80 million) and ChinaMAP (7.98 million) reference panels. When focusing on only high-quality imputed variants, we observed that the ChinaMAP reference panel had the greatest percentage of high-quality variants (86.23%), followed by CKB (84.63%), TOPMed (80.22%), extended 1KGP (78.50%) and NyuWa (77.32%) (Figure 3a, Table 2). We note that the reason why the ChinaMAP provides the smallest number of medium-quality variants is that it automatically filters out almost half of low-quality variants in the actually used panel.

We evaluated the imputation accuracy by using three measurements: Pearson correlation coefficient (R^2), precision and sensitivity. The mean R^2 of the compared reference panels were 0.964 (ChinaMAP), 0.961 (CKB), 0.946 (TOPMed), 0.943 (NyuWa) and 0.926 (extended 1KGP) (Figure 3b). For the ratios of true positive (TP), false positive (FP) and false negative (FN) variants, the ChinaMAP reached the highest ratio of TP variants (94.62%), subsequently followed by CKB (93.71%), then followed by TOPMed (92.21%), NyuWa (92.18%) and extended 1KGP (90.09%). Meanwhile, the ChinaMAP obtained the lowest ratios of FP (1.71%) and FN (3.68%) variants, and for the CKB panel, the two ratios were 1.93 and 4.35%, respectively. These ratios in TOPMed (FP: 2.35 and FN: 5.43%) and NyuWa (FP: 2.57% and FN: 5.28%) were slightly higher than those in the CKB panel. The extended 1KGP reference panel had the highest ratio of FP (3.08%) and FN (6.85%) variants (Figure 3c). Consequently,

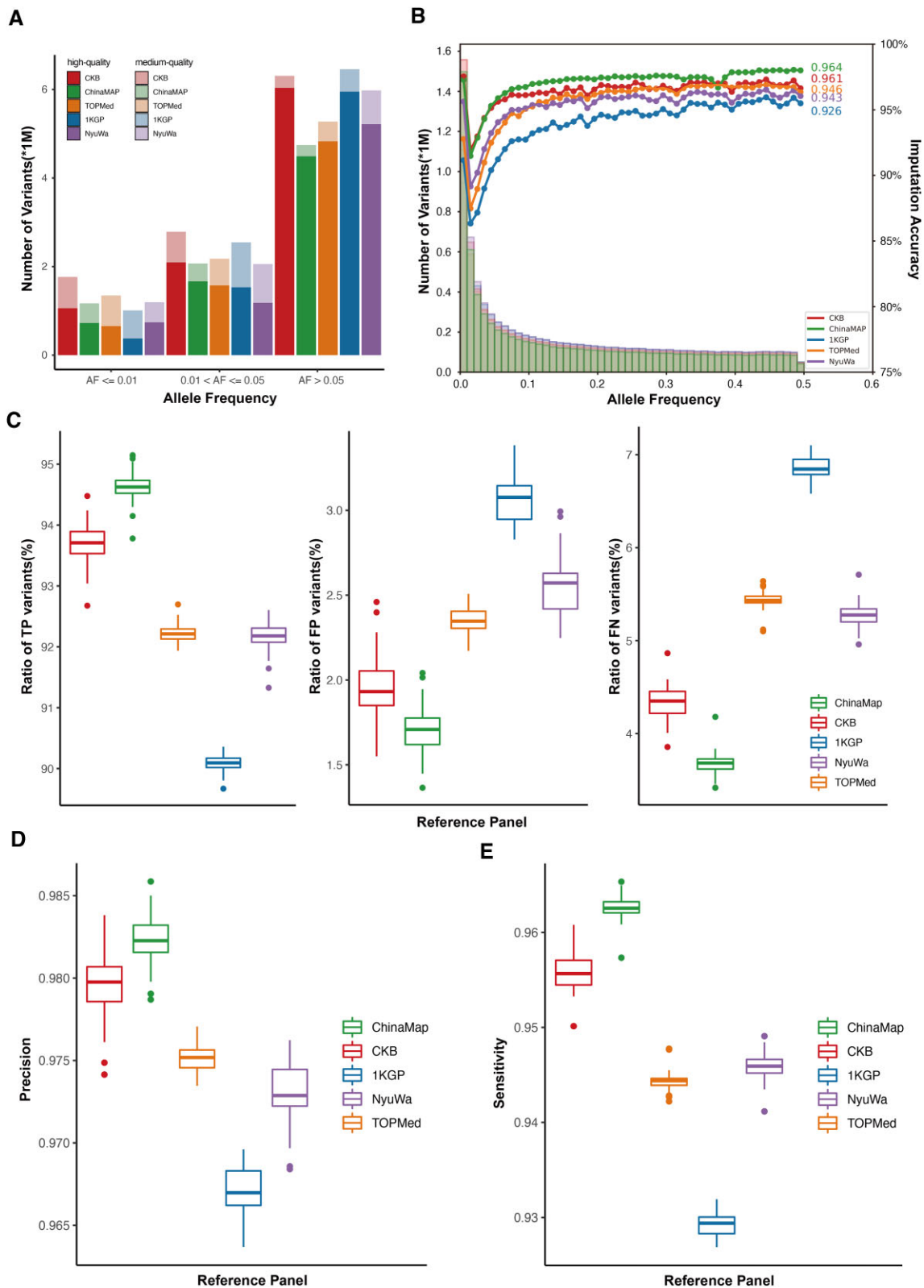


Figure 3. The performance for imputing 50 CKB microarray data. **(A)** The numbers of high- and medium-quality imputed variants under different AF (allele frequency) by using different reference panels. **(B)** The histogram of imputed variants and Pearson correlation coefficients for different panels. **(C)** The boxplots of the ratios of true positive (TP), false negative (FN) and false positive (FP) variants. **(D)** The imputation precision of reference panels. **(E)** The sensitivity of the reference panels.

Table 2. The high-quality and medium-quality imputed variants for imputing 50 microarray samples

| Reference panels | Type | AF \leq 1% | 1% < AF \leq 5% | AF > 5% | ALL |
|------------------|--------------------|--------------|-------------------|---------|--------|
| CKB | medium-quality (M) | 0.71 | 0.69 | 0.27 | 1.67 |
| | high-quality (M) | 1.06 | 2.10 | 6.04 | 9.19 |
| | high-quality rate | 0.6000 | 0.7519 | 0.9570 | 0.8463 |
| ChinaMAP | medium-quality (M) | 0.44 | 0.40 | 0.25 | 1.10 |
| | high-quality (M) | 0.73 | 1.67 | 4.49 | 6.88 |
| | high-quality rate | 0.6205 | 0.8052 | 0.9468 | 0.8623 |
| 1KGP | medium-quality (M) | 0.63 | 1.01 | 0.51 | 2.15 |
| | high-quality (M) | 0.38 | 1.53 | 5.95 | 7.86 |
| | high-quality rate | 0.3730 | 0.6021 | 0.9217 | 0.7850 |
| TOPMed | medium-quality (M) | 0.69 | 0.60 | 0.45 | 1.74 |
| | high-quality (M) | 0.66 | 1.58 | 4.83 | 7.06 |
| | high-quality rate | 0.4867 | 0.7234 | 0.9154 | 0.8022 |
| NyuWa | medium-quality (M) | 0.45 | 0.88 | 0.76 | 2.09 |
| | high-quality (M) | 0.74 | 1.18 | 5.22 | 7.14 |
| | high-quality rate | 0.6202 | 0.5741 | 0.8723 | 0.7732 |

Notes: (M) represents million.

the ChinaMAP attained the highest precision of 98.23%, followed by CKB (97.98%), TOPMed (97.52%), NyuWa (97.29%) and extended 1KGP (96.70%). For sensitivity, the ChinaMAP and CKB panels reached 96.25 and 95.57%, respectively. Following that, the NyuWa, TOPMed and extended 1KGP obtained sensitivities of 94.59, 94.44 and 92.94%, respectively. The CKB reference panel achieved very similar R^2 , precision and sensitivity compared to the ChinaMAP, displaying an outstanding imputation performance (Figure 3d and e).

In the imputation of the 1KGP array data, we compared the performance of the CKB panel with that of ChinaMAP and TOPMed. We excluded the extended 1KGP panel as it had overlap samples with the array data, and also excluded NyuWa panel as the web server is unstable and not accessible for submitting jobs currently. The CKB reference panel provided the highest number of medium-quality imputed variants (9.75 million), followed by TOPMed (7.83 million) and ChinaMAP (6.98 million) reference panels. When focusing on only high-quality imputed variants, we observed that the ChinaMAP reference panel had the greatest percentage of high-quality imputed variants (87.92%), followed by TOPMed (84.46%) and CKB (84.11%) (Supplementary Figure S3a). For the Pearson correlation coefficient R^2 , both the CKB and ChinaMAP panels achieved 0.979, while the TOPMed had a lower R^2 of 0.965 (Supplementary Figure S3b).

Imputation of 100 706 microarray data

For the 100 706 samples with microarray data, we provided their sex and age distribution in each sampling site (Figure 4a). Specifically, the provinces of Heilongjiang ($N = 13\ 131$), Hunan ($N = 12\ 512$), Zhejiang ($N = 12\ 042$), Henan ($N = 11\ 421$), Sichuan ($N = 10\ 637$) and Gansu ($N = 10\ 058$) had recruitments >10 000. The province of Hainan had smallest recruitment of 5794. The PCA of microarray data before imputation was provided in Figure 4b. The PC1 represents the latitudinal gradient. The imputation-completed whole genome data contained 42.61 million medium-quality variants and 17.45 million high-quality variants. To assess the imputation performance of the 100 706 CKB microarray data, we calculated the Pearson correlation coefficients (R^2) of 50 CKB samples with imputed genotype and high-depth WGS data. Note that we did not have WGS data for the 100 706 samples, thus

we could not use that as the true set. As an alternative, we used a subset of 50 individuals with WGS data as samples being evaluated. Consequently, the averaged R^2 was 0.972. Remember that when we simulated only these 50 microarray samples, the averaged R^2 was 0.961 (Supplementary Figure S4). This R^2 difference may be due to the randomness of the imputation algorithm in Beagle v.5.2, hidden Markov model (28). The high imputation accuracy of 0.972 demonstrated that the proposed CKB reference panel is quite capable of imputing extensive data.

GWAS analysis of simulated data

With imputed phenotype data under the null hypothesis that there were no associated SNPs, the GWAS analysis did not identify any significant signals and the P -values were uniformly distributed (Supplementary Figure S5a and b) as expected. When the phenotype was generated by involving the effects of SNPs, the GWAS study successfully discovered causal SNPs and those in high linkage disequilibrium (LD) (Supplementary Figure S5c). Specifically, in addition to the five randomly selected causal SNPs (rs3003378, rs6764623, rs10905649, rs13254191 and rs10915307), high-LD SNPs (e.g. rs12564681, rs11923809, rs7092291, rs545854, rs12123277) were also identified. The results of GWAS analysis with simulated data under both null and alternative hypotheses demonstrated the high-quality of genotype data.

GWAS analysis of real phenotype data

After filtering in SNPs with $MAF > 0.01$, HWE P -value $> 1E-06$, and genotype missing rate < 0.01 , the numbers of SNPs in GWAS analysis before and after imputation were 3 038 178 and 9 205 896, respectively. The increase in number of SNPs was substantial. At the significance threshold of $5E-08$, the number of significant SNPs increased from 7971 to 16 508 after imputation (Figure 5, Supplementary Table S4). The numbers of identified significant loci for original and after-imputed data were 119 and 147, respectively. The shared 119 loci included the well-known height-associated genes *GDF5* (cartilage-derived morphogenetic protein 1) (48), *IGF1R* (insulin-like growth factor 1 receptor) (49), and *ADCY3* (ATP pyrophosphate-Lyase 3) (50). Among the additional 28 loci, 26 (92.9%) were previously reported to be associated with

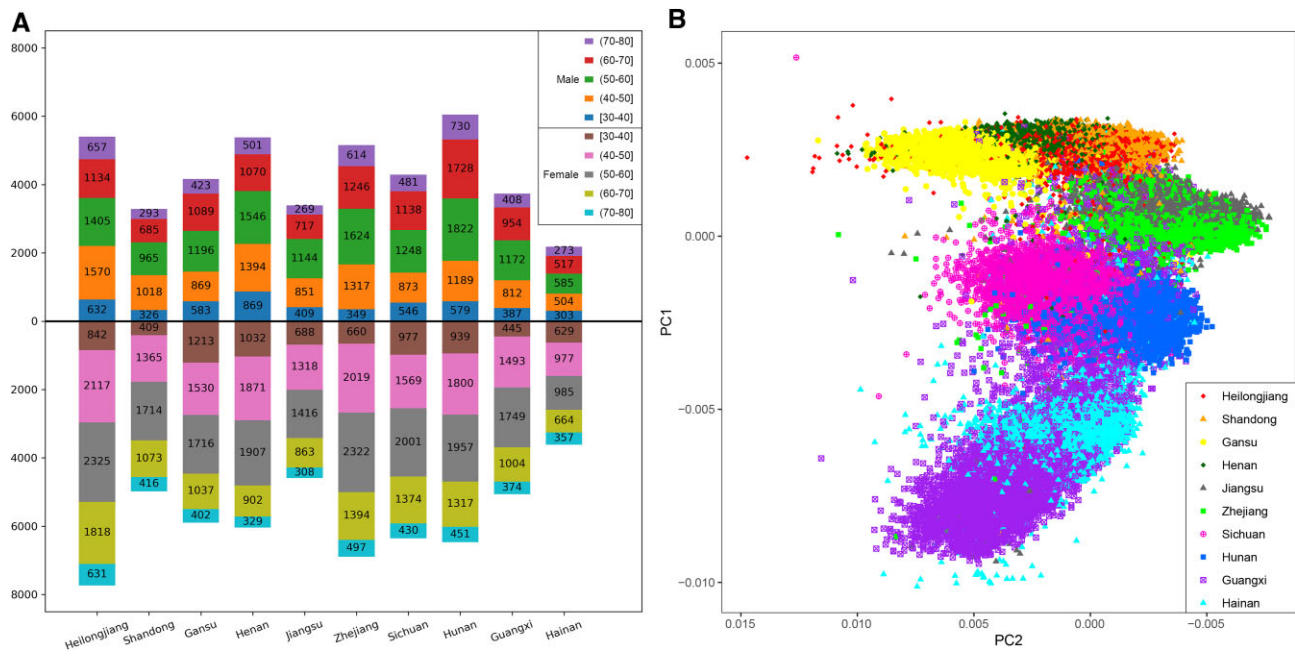


Figure 4. The sample information and principal component analysis of the microarray data. **(A)** The sex and age distribution of samples in each sampling site. The age distributions of males (females) were on the top (bottom) of the x-axis. The total number of samples from all sampling sites was 100 640, as 66 samples with missing sampling site information. **(B)** The principal component analysis of 100 706 samples with microarray data before genotype imputation. The PC1 represents a latitudinal gradient, from north to south China. Each color represents a province of sampling site.

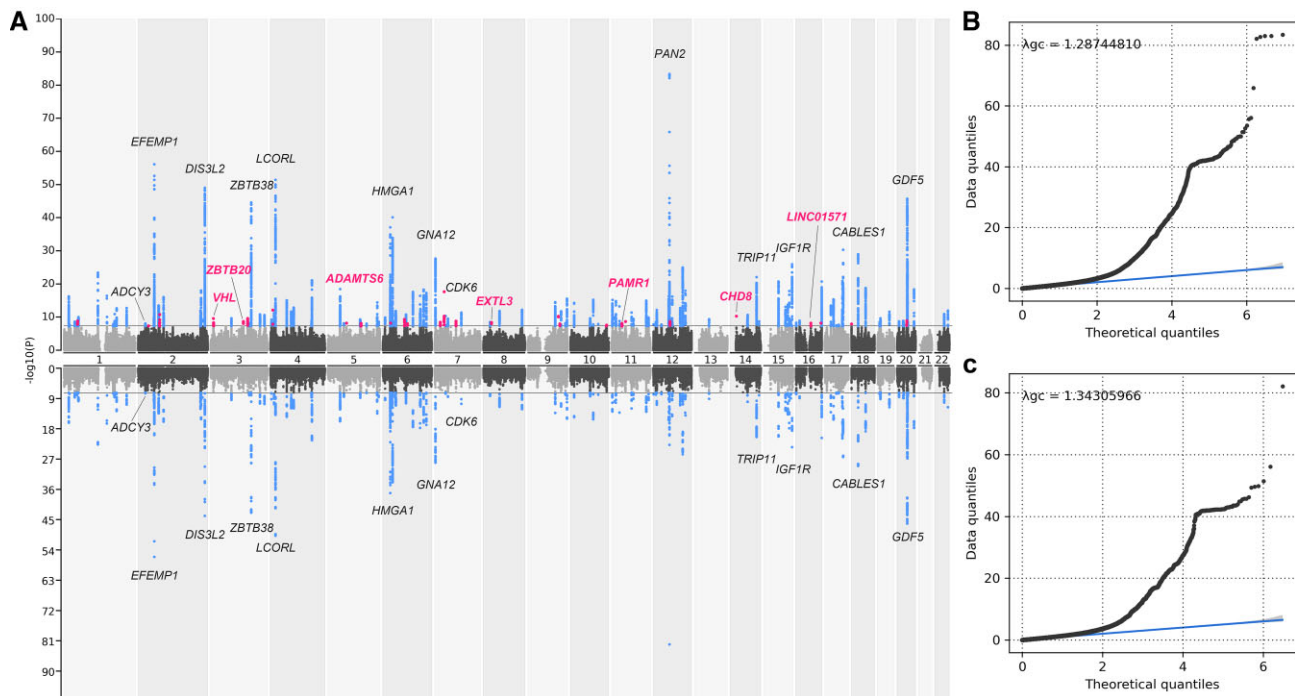


Figure 5. The GWAS results of height before and after genotype imputation. **(A)** The mirrored Manhattan plots of GWAS results based on the microarray data after (top) and before (bottom) genotype imputation. Genes in black are a list of shared genes identified before and after imputation. Genes in purple are a list of representative genes identified only after imputation. **(B)** The QQ-plot of GWAS results after genotype imputation. **(C)** The QQ-plot of GWAS results before genotype imputation. The genomic inflation factors (λ_{gc}) were 1.287 and 1.343 after and before imputation, respectively.

height, for example *CHD8* (chromodomain helicase DNA binding protein 8) functioned in transcriptional regulation and promotion of cell proliferation (51), *ZBTB20* (zinc finger protein 288) played a role in glucose homeostasis and post-natal growth (51), and *PAMR1* (regeneration-associated muscle protease homolog) might played a role in regeneration of skeletal muscle (51). The GWAS results with real phenotype indicated the high quality and credibility of the imputed data.

Discussion

A population-specific haplotype reference panel is a collection of ancestral chromosome sequences that represents the genetic diversity of the population. A high-precision reference panel is the basis for population genetic analysis and precision medicine. China has the largest population in the world and possesses vast amounts of genetic resources, but lacks a high-quality reference panel, which has hindered the development of genetic studies and their application in human diseases based on the Chinese population. Fortunately, in the last 2 years, a few reference panels have been constructed for accurate genotype imputation in the Chinese population, including the ChinaMAP and NyuWa.

In this work, we developed a high-resolution haplotype-resolved reference panel of 10 000 sequenced individuals from the CKB cohort and the 1KGP database. Even with medium sequencing depth (15.41×), the proposed CKB panel can compete with the ChinaMAP (40.80×) and outperform the extended 1KGP, TOPMed and NyuWa in imputation accuracy measured by Pearson correlation coefficient, precision and sensitivity. From the perspective of the number of well-imputed variants, the CKB provided the largest number of medium-quality variants with an information score between 0.4 and 0.8; for high-quality variants with an information score >0.8, the CKB panel obtained the second largest amount among all considered panels. What is more valuable is that we completed the genotype imputation for 100 706 CKB microarray data based on the constructed panel. The imputation accuracy reached as high as 0.972 and GWAS analysis based on the simulated data and the real phenotype height demonstrated the reliability of the extensive imputed data. This imputed dataset is the largest whole genome data for Chinese population to date and will certainly play a fundamental role in personalized medicine and drug development.

However, it must be acknowledged that our study has some limitations. First, the sequencing depth is medium (~15×). Based on our evaluation, compared to high coverage data (>30×), medium sequencing data have comparable base quality measured by Q20, Q30 and GC content. However, the genomic coverage at different sequencing depth has differences, especially for higher coverage. In detail, for 1×, 4× and 10×, the coverage differences are about 0.2, 1.0 and 18%, respectively, which might have influence on rare and novel variants detection. We note that the comparison results were obtained from two particular datasets and could not represent a general tendency. Second, 9914 out of 9964 (99.50%) subjects in the CKB reference panel were stroke cases, even though the results of variants detection and association analysis were promising, the explicit influence of potential disease haplotype is hard to tell and needs further investigation.

The ultimate goal of imputing genotype data is to increase statistical power of genetic association studies for identifying trait-associated SNPs and to reveal the etiology of com-

plex diseases. As the hitherto largest cohort of Chinese population, CKB collected abundant clinical data, including demographic, anthropometric, biochemical, radiographic traits, metabolomic tests and diseases coded by ICD10 (international classification of diseases, v.10). There are >1500 diseases, mostly chronic, such as heart attack, stroke, diabetes, cancers and so on. As a significant future work, we aim to perform GWAS analysis for the vast wealth of phenotypes and over 100 000 imputed WGS genotype data. In recent years, as a precision medicine tool, the polygenic risk score, also known as the polygenic score, has been widely used to predict an individual's genetic risk of disease. The predictive accuracy of the polygenic risk score largely relies on the sample sizes in discovery samples. To the best of our knowledge, with the after-imputed genomic data, it should be the largest population genetic study of the Chinese population and is also comparable to numerous international genomics research projects, for example, the UK Biobank study (<https://www.ukbiobank.ac.uk/>), the All of Us research program (<https://allofus.nih.gov/>) and the biobank Japan project (<https://biobank.jp.org/en/>).

Most of the reference panels are now packaged into online imputation servers, such as the Michigan imputation server (40), TOPMed imputation server (40), ChinaMAP imputation server, NyuWa server and our developed CKB imputation server. These imputation servers all provide free genotype imputation service by uploading to-be-imputed files and selecting reference panel, population and imputation software. All the imputation results can be downloaded directly by clicking on filenames. Even though the online server provides a convenient way to impute genotype data, it typically cannot handle large-sized files, which causes difficulties in imputing large-sample data. When imputing large-scale datasets, the individual-level reference panels are needed for offline imputation. Since the completion of the first human genome project in 2003 (<https://www.genome.gov/human-genome-project>), the only database that is fully publicly available is the 1000 Genomes Database. Sharing genomic data is critical for research efficiency, translating research results into clinical applications and ultimately improving public health. Hence, we appeal for the sharing of genomic and health-related data with controlled management.

Data availability

The CKB reference panel and the after-imputed >100 000 CKB microarray data have been deposited into CNGB Sequence Archive (CNSA) of China National GeneBank DataBase (CNGBdb) with accession number CNP0003405. All genotype data are shared with controlled management.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

The most important acknowledgement is to the participants in the study and the members of the survey teams in each of the 10 regional centers, as well as to the project development and management teams based at Beijing, Oxford and the 10 regional centers.

Author contribution: L.L., X.J., J.L. and H.Z. conceived the study, designed the research program and managed the project. C.Y., Y.G., D.S., Z.Y., L.D., F.R. and P.Q. finished the laboratory processing and data acquisition. C.Y., X.L., Y.T., Y.G., D.S., P.Q. and Y.Z. preprocessed the data, finished the quality control and constructed the haplotype reference panel. Y.T., X.L., L.L., P.P., J.Z., R.G. and X.J. finished the evaluation of the haplotype reference panel. R.W., I.M., H.D., L.Y., Y.C. and Z.C. provided useful advice on constructing and evaluating the reference panel. T.Y., F.Y., F.C. and X.W. built the online imputation server. X.L., Y.T. and H.Z. drafted the manuscript. C.Y., X.L., Y.T., L.L., Y.Z., H.Z. and X.J. mainly organized the revised version and performed data analysis. All authors participated in revising the manuscript.

Funding

This work was supported by grants National Natural Science Foundation of China (32000398, 82192901, 82192904, 82192900), the National Key R&D Program of China (2016YFC0900500), the China National GeneBank, Guangdong Provincial Key Laboratory of Genome Read and Write (2017B030301011), Guangdong Provincial Academician Workstation of BGI Synthetic Genomics (2017B090904014), open project of BGI-Shenzhen, Shenzhen 518000 China (BGIRSZ20200008) and the Innovation Platform for Academicians of Hainan Province (YSPTZX202118). The CKB baseline survey and the first re-survey were supported by a grant from the Kadoorie Charitable Foundation in Hong Kong. The long-term follow-up is supported by grants from the UK Wellcome Trust (212946/Z/18/Z, 202922/Z/16/Z, 104085/Z/14/Z, 088158/Z/09/Z), National Natural Science Foundation of China (81390540, 91846303, 81941018) and Chinese Ministry of Science and Technology (2011BAI09B01). The funders had no role in the study design, data collection, data analysis and interpretation, writing of the report or the decision to submit the article for publication.

Conflict of interest statement

None declared.

References

- Dugger,S.A., Platt,A. and Goldstein,D.B. (2018) Drug development in the era of precision medicine. *Nat. Rev. Drug Discovery*, **17**, 183–196.
- Gough,A., Soto-Gutierrez,A., Vernetti,L., Ebrahimkhani,M.R., Stern,A.M. and Taylor,D. (2021) Human biomimetic liver microphysiology systems in drug development and precision medicine. *Nat. Rev. Gastroenterol. Hepatol.*, **18**, 252–268.
- International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, **437**, 1299.
- International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851.
- Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature*, **526**, 68.
- Byrska-Bishop,M., Evani,U.S., Zhao,X., Basile,A.O., Abel,H.J., Regier,A.A., Corvelo,A., Clarke,W.E., Musunuri,R. and Nagulapalli,K. (2022) High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*, **185**, 3426–3440.
- McCarthy,S., Das,S., Kretzschmar,W., Delaneau,O., Wood,A.R., Teumer,A., Kang,H.M., Fuchsberger,C., Danecek,P. and Sharp,K. (2016) A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.*, **48**, 1279.
- Taliun,D., Harris,D.N., Kessler,M.D., Carlson,J., Szpiech,Z.A., Torres,R., Taliun,S.A.G., Corvelo,A., Gogarten,S.M. and Kang,H.M. (2021) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, **590**, 290–299.
- Francioli,L.C., Menelaou,A., Pulit,S.L., Van Dijk,F., Palamara,P.F., Elbers,C.C., Neerincx,P.B., Ye,K., Guryev,V. and Kloosterman,W.P. (2014) Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.*, **46**, 818–825.
- Marett,L., Jensen,J.M., Petersen,B., Sibbesen,J.A., Liu,S., Villesen,P., Skov,L., Belling,K., Theil Have,C. and Izarzugaza,J.M. (2017) Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. *Nature*, **548**, 87–91.
- Gudbjartsson,D.F., Helgason,H., Gudjonsson,S.A., Zink,F., Oddson,A., Gylfason,A., Besenbacher,S., Magnusson,G., Halldorsson,B.V. and Hjartarson,E. (2015) Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.*, **47**, 435–444.
- Wu,D., Dou,J., Chai,X., Bellis,C., Wilm,A., Shih,C.C., Soon,W.W.J., Bertin,N., Lin,C.B. and Khor,C.C. (2019) Large-scale whole-genome sequencing of three diverse Asian populations in Singapore. *Cell*, **179**, 736–749.
- Cao,Y., Li,L., Xu,M., Feng,Z., Sun,X., Lu,J., Xu,Y., Du,P., Wang,T. and Hu,R. (2020) The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. *Cell Res.*, **30**, 717–731.
- Li,L., Huang,P., Sun,X., Wang,S., Xu,M., Liu,S., Feng,Z., Zhang,Q., Wang,X. and Zheng,X. (2021) The ChinaMAP reference panel for the accurate genotype imputation in Chinese populations. *Cell Res.*, **31**, 1308–1310.
- Zhang,P., Luo,H., Li,Y., Wang,Y., Wang,J., Zheng,Y., Niu,Y., Shi,Y., Zhou,H. and Song,T. (2021) NyuWa Genome resource: a deep whole-genome sequencing-based variation profile and reference panel for the Chinese population. *Cell Rep.*, **37**, 110017.
- Chen,Z., Lee,L., Chen,J., Collins,R., Wu,F., Guo,Y., Linksted,P. and Peto,R. (2005) Cohort profile: the Kadoorie study of chronic disease in China (KSCDC). *Int. J. Epidemiol.*, **34**, 1243–1249.
- Chen,Z., Peto,R., Zhou,M., Iona,A., Smith,M., Yang,L., Guo,Y., Chen,Y., Bian,Z. and Lancaster,G. (2015) Contrasting male and female trends in tobacco-attributed mortality in China: evidence from successive nationwide prospective cohort studies. *Lancet North Am. Ed.*, **386**, 1447–1456.
- Bennett,D.A., Du,H., Clarke,R., Guo,Y., Yang,L., Bian,Z., Chen,Y., Millwood,I., Yu,C. and He,P. (2017) Association of physical activity with risk of major cardiovascular diseases in Chinese men and women. *JAMA Cardiol.*, **2**, 1349–1358.
- Du,H., Li,L., Bennett,D., Guo,Y., Key,T.J., Bian,Z., Sherliker,P., Gao,H., Chen,Y. and Yang,L. (2016) Fresh fruit consumption and major cardiovascular disease in China. *N. Engl. J. Med.*, **374**, 1332–1343.
- Qin,C., Lv,J., Guo,Y., Bian,Z., Si,J., Yang,L., Chen,Y., Zhou,Y., Zhang,H. and Liu,J. (2018) Associations of egg consumption with cardiovascular disease in a cohort study of 0.5 million Chinese adults. *Heart*, **104**, 1756–1763.
- Bragg,F., Holmes,M.V., Iona,A., Guo,Y., Du,H., Chen,Y., Bian,Z., Yang,L., Herrington,W. and Bennett,D. (2017) Association between diabetes and cause-specific mortality in rural and urban areas of China. *JAMA*, **317**, 280–289.
- Yu,C., Tang,H., Guo,Y., Bian,Z., Yang,L., Chen,Y., Tang,A., Zhou,X., Yang,X. and Chen,J. (2018) Hot tea consumption and its interactions with alcohol and tobacco use on the risk for esophageal cancer: a population-based cohort study. *Ann. Intern. Med.*, **168**, 489–497.
- Walters,R.G., Millwood,I.Y., Lin,K., Schmidt Valle,D., McDonnell,P., Hacker,A., Avery,D., Edris,A., Fry,H., Cai,N., et al.

- (2023) Genotyping and population characteristics of the China Kadoorie Biobank. *Cell Genom.*, **3**, 100361.
24. Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., Li, Y., Ye, J., Yu, C. and Li, Z. (2018) SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Gigascience*, **7**, gix120.
 25. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics*, **43**, 11.10.1–11.10.33.
 26. Jun, G., Flickinger, M., Hetrick, K.N., Romm, J.M., Doheny, K.F., Abecasis, G.R., Boehnke, M. and Kang, H.M. (2012) Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. Hum. Genet.*, **91**, 839–848.
 27. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A. and Hanna, M. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.
 28. Browning, B.L., Tian, X., Zhou, Y. and Browning, S.R. (2021) Fast two-stage phasing of large-scale sequence data. *Am. Hum. Genet.*, **108**, 1880–1890.
 29. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M. and Chen, W.-M. (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics*, **26**, 2867–2873.
 30. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P. and Cunningham, F. (2016) The ensembl variant effect predictor. *Genome Biol.*, **17**, 122.
 31. Ng, P.C. and Henikoff, S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.
 32. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. and Sunyaev, S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
 33. Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M. and Maglott, D.R. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, **42**, D980–D985.
 34. Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D. and Hoover, J. (2016) ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.*, **44**, D862–D868.
 35. Novembre, J. and Stephens, M. (2008) Interpreting principal component analyses of spatial population genetic variation. *Nat. Genet.*, **40**, 646–649.
 36. Patterson, N., Price, A.L. and Reich, D. (2006) Population structure and eigenanalysis. *PLoS Genet.*, **2**, e190.
 37. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I. and Daly, M.J. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. Hum. Genet.*, **81**, 559–575.
 38. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M. and Lee, J.J. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, **4**, 7.
 39. Fuchsberger, C., Abecasis, G.R. and Hinds, D.A. (2015) minimac2: faster genotype imputation. *Bioinformatics*, **31**, 782–784.
 40. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S. and McGue, M. (2016) Next-generation genotype imputation service and methods. *Nat. Genet.*, **48**, 1284–1287.
 41. Browning, B.L., Zhou, Y. and Browning, S.R. (2018) A one-penny imputed genome from next-generation reference panels. *Am. Hum. Genet.*, **103**, 338–348.
 42. Wang, J., Raskin, L., Samuels, D.C., Shyr, Y. and Guo, Y. (2015) Genome measures used for quality control are dependent on gene function and ancestry. *Bioinformatics*, **31**, 318–323.
 43. Sherry, S.T., Ward, M. and Sirotkin, K. (1999) dbSNP—database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.*, **9**, 677–679.
 44. McCarthy, D.J., Humburg, P., Kanapin, A., Rivas, M.A., Gaulton, K., Cazier, J.-B. and Donnelly, P. (2014) Choice of transcripts and software has a large effect on variant annotation. *Genome Medicine*, **6**, 26.
 45. Tan, A., Abecasis, G.R. and Kang, H.M. (2015) Unified representation of genetic variants. *Bioinformatics*, **31**, 2202–2204.
 46. Kowalski, M., Qian, H., Hou, Z., Rosen, J., Tapia, A., Shan, Y., Jain, D., Argos, M., Arnett, D. and Avery, C. (2019) NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium; TOPMed Hematology & Hemostasis Working Group: use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genet.*, **15**, e1008500.
 47. Zhang, P., Luo, H., Li, Y., Wang, Y., Wang, J., Zheng, Y., Niu, Y., Shi, Y., Zhou, H., Song, T., et al. (2021) NyuWa Genome resource: a deep whole-genome sequencing-based variation profile and reference panel for the Chinese population. *Cell Rep.*, **37**, 110017.
 48. Sanna, S., Jackson, A.U., Nagaraja, R., Willer, C.J., Chen, W.M., Bonnycastle, L.L., Shen, H., Timpson, N., Lettre, G., Usala, G., et al. (2008) Common variants in the GDF5-UQC region are associated with variation in human height. *Nat. Genet.*, **40**, 198–203.
 49. Fontenele, E.G., Moraes, M.E., d’Alva, C.B., Pinheiro, D.P., Landim, S.A., Barros, F.A., Trarbach, E.B., Mendonca, B.B. and Jorge, A.A. (2015) Association study of GWAS-derived loci with height in Brazilian children: importance of MAP3K3, MMP24 and IGF1R polymorphisms for height variation. *Horm Res Paediatr.*, **84**, 248–253.
 50. Stergiakouli, E., Gaillard, R., Tavaré, J.M., Balthasar, N., Loos, R.J., Taal, H.R., Evans, D.M., Rivadeneira, F., St Pourcain, B., Uitterlinden, A.G., et al. (2014) Genome-wide association study of height-adjusted BMI in childhood identifies functional variant in ADCY3. *Obesity (Silver Spring)*, **22**, 2252–2259.
 51. Yengo, L., Vedantam, S., Marouli, E., Sidorenko, J., Bartell, E., Sakaue, S., Graff, M., Eliassen, A.U., Jiang, Y., Raghavan, S., et al. (2022) A saturated map of common genetic variants associated with human height. *Nature*, **610**, 704–712.

Original Paper

Immune-Boosting Effect of the COVID-19 Vaccine: Real-World Bidirectional Cohort Study

Ming Liu^{1*}, MM; Tianshuo Zhao^{2,3,4,5*}, MD; Qiuyue Mu¹, MPH; Ruizhi Zhang¹, MPH; Chunting Liu¹, MM; Fei Xu¹, MPH; Luxiang Liang¹, AD; Linglu Zhao¹, MM; Suye Zhao¹, BMed; Xianming Cai^{2,3,4,5}, MM; Mingting Wang^{2,3,4,5}, MPH; Ninghua Huang^{2,3,4,5}, MD; Tian Feng¹, MPH; Shiguang Lei¹, BMed; Guanghong Yang¹, MD; Fuqiang Cui^{2,3,4,5}, MD, MPH, MPM, PhD

¹Guizhou Center for Disease Control and Prevention, Guiyang, China

²Department of Laboratorial Science and Technology, School of Public Health, Peking University, Beijing, China

³Vaccine Research Center, School of Public Health, Peking University, Beijing, China

⁴Center for Infectious Diseases and Policy Research & Global Health and Infectious Diseases Group, Peking University, Beijing, China

⁵Key Laboratory of Epidemiology of Major Diseases, Peking University, Ministry of Education, Beijing, China

* these authors contributed equally

Corresponding Author:

Guanghong Yang, MD

Guizhou Center for Disease Control and Prevention

73 Bageyan Road

Yunyan District, Guizhou

Guiyang, 550004

China

Phone: 86 0851 86828805

Email: ghyang_gzmu@outlook.com

Abstract

Background: As the SARS-CoV-2 attenuates and antibodies from the COVID-19 vaccine decline, long-term attention should be paid to the durability of primary booster administration and the preventive effect of the second or multiple booster doses of the COVID-19 vaccine.

Objective: This study aimed to explore the durability of primary booster administration and the preventive effect of second or multiple booster doses of the COVID-19 vaccine.

Methods: We established a bidirectional cohort in Guizhou Province, China. Eligible participants who had received the primary booster dose were enrolled for blood sample collection and administration of the second booster dose. A retrospective cohort for the time of administration was constructed to evaluate antibody attenuation 6-12 months after the primary booster dose, while a prospective cohort on the vaccine effect of the second booster dose was constructed for 4 months after the second administration.

Results: Between September 21, 2022, and January 30, 2023, a total of 327 participants were included in the final statistical analysis plan. The retrospective cohort revealed that approximately 6-12 months after receiving the primary booster, immunoglobulin G (IgG) slowly declined with time, while immunoglobulin A (IgA) remained almost constant. The prospective cohort showed that 28 days after receiving the second booster, the antibody levels were significantly improved. Higher levels of IgG and IgA were associated with better protection against COVID-19 infection for vaccine recipients. Regarding the protection of antibody levels against post-COVID-19 symptoms, the increase of the IgG had a protective effect on brain fog and sleep quality, while IgA had a protective effect on shortness of breath, brain fog, impaired coordination, and physical pain.

Conclusions: The IgG and IgA produced by the second booster dose of COVID-19 vaccines can protect against SARS-CoV-2 infection and may alleviate some post-COVID-19 symptoms. Further data and studies on secondary booster administration are required to confirm these conclusions.

(*JMIR Public Health Surveill* 2023;9:e47272) doi: [10.2196/47272](https://doi.org/10.2196/47272)

KEYWORDS

bidirectional cohort study; booster administration; COVID-19 vaccine; real-world study; SARS-CoV-2; vaccine efficacy; COVID-19

Introduction

Since the end of 2019, COVID-19 has been the cause of a global pandemic, placing a heavy burden on the global public health system [1]. With the widespread and continuous evolution of SARS-CoV-2, many variants of concern (VOCs), such as Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1), Delta (B.1.617.2), and Omicron (B.1.1.529), have emerged globally and have led to several infection waves [2-5]. At present, the Omicron variant, which has a higher transmissibility and immune escape ability, is the dominant variant in the world [6]. Previous studies have shown that the Omicron variant not only has resistance to serum antibodies of convalescent patients but also has certain resistance to the serum of individuals who have been fully vaccinated against COVID-19 [7-12]. Therefore, Omicron poses a serious threat to the control of the COVID-19 pandemic and disease treatment.

China has administered 3.491 billion doses of the COVID-19 vaccine. The coverage rate of the first dose and second dose for the entire population reached 92.9% and 90.6%, respectively [13], and more than 771 million booster injections have been administered [14]. In November 2022, China adjusted and optimized the prevention and control measures for COVID-19. The local epidemic quickly climaxed, which led to a depletion of medical resources. Febrifuge, antitussive, and other COVID-19-related drugs could not meet the exponential increase in the number of patients in the short term, and the more serious concern was overwhelming the availability of hospital beds [15,16]. Therefore, long-term attention should be paid to the preventive effect and clinical value of the second or multiple booster doses of the COVID-19 vaccine.

In China, the most commonly used vaccines for both primary and booster immunization against COVID-19 are inactivated vaccines produced by the China National Biotec Group and Sinovac Biotech. Inactivated vaccines are prepared by cultivating SARS-CoV-2 in vitro to render the virus noninfectious while preserving its antigenicity. Although homologous boosting is generally considered a standard practice, heterologous regimens have been proposed as a COVID-19 vaccine strategy to elicit stronger and broader, or longer-lasting, immunity [17,18]. A recombinant COVID-19 vaccine using adenovirus type 5 as a vector for inhalation was developed by CanSino Biologics (inhalant Ad5-nCoV). Ad5-nCoV inhalation involves the recombination of the spike glycoprotein gene of SARS-CoV-2 into the replication-deficient human type 5 adenovirus gene, which induces an immune response in the body. This inhalant is easy to administer and can stimulate mucosal immunity. In a previous phase 1 trial, Ad5-nCoV inhalation was found to be well tolerated. Further, compared with intramuscular vaccination, aerosol vaccination could trigger a higher ratio of neutralizing antibodies to total antibodies [19].

Few real-world studies have demonstrated the effect of a fourth dose of heterologous booster on Omicron, especially using

inhalation vaccines. Here, we aimed to reveal the immunogenicity and persistence of the primary booster dose and the real-world immune effect of the secondary inhalation booster to assess immunogenicity and persistence and prevent sequelae of booster administration in the real world.

Methods

Study Design

Overview

A bidirectional cohort was established to investigate the efficacy of booster administration between September 21, 2022, and January 30, 2023, in Guiyang City, Guizhou Province, China. The cohort was retrospectively tracked to determine the effect and durability of primary booster administration and prospectively followed up to assess the immunogenicity and real-world protective effect of secondary booster heterologous immunization, with the time of receiving the second booster as the node.

Retrospective Study

Blood samples were collected from individuals for antibody testing at the time of enrollment. A questionnaire covering basic characteristics and immunization programs was required to evaluate the durability and effectiveness of the third booster dose.

Secondary Booster Administration

All enrolled individuals received the inhalant, Ad5-nCoV, as part of a secondary booster immunization program (the fourth dose).

Prospective Study

About 4 weeks (21-35 days) after receiving the second booster dose, blood samples were collected for antibody testing. After about 16 weeks (84-140 days), information on the infection and sequelae of COVID-19 was collected from participants through follow-up phone calls.

Participants

Participants were recruited by the Guizhou Center for Disease Control and Prevention. These individuals were aged 18 years or older and had received 3 doses of the COVID-19 vaccine before 6 months or above. The main exclusion criteria were individuals with a history of clinically or laboratory-confirmed COVID-19 or SARS-CoV-2 infection within the first 6 months of enrollment, a history of vaccination (any administration, including COVID-19 baseline or booster dose) within the first 6 months of enrollment, or an allergy to any component of the vaccine.

Procedures

Individuals from Guanshanhu District, Qingzhen City, and Baiyun District of Guiyang City were recruited for this study. All individuals completed the basic and primary booster

immunization procedures with the inactivated vaccine from the China National Biotec Group or Sinovac Biotech.

Eligible participants received 1 dose of inhalant Ad5-nCoV (0.1 mL per dose) through a specific atomization device. Venous blood samples (5 mL) were collected before inhalation and 28 days after inhalation to detect immunoglobulin G (IgG) and immunoglobulin A (IgA) antibodies against SARS-CoV-2 in serum. Antibody detection was performed by the receptor-binding domain antibody test kit produced by Vazyme Biotech Co Ltd. The kit detects receptor-binding domains IgA and IgG antibodies against SARS-CoV-2 that are produced during incubation through an indirect enzyme-linked immunosorbent assay (ELISA). After processing and color development, the absorbance of the sample was measured at a wavelength of 450 nm. The absorbance of the sample was positively correlated with the antibody titers.

Survey Tool

Telephone follow-up was conducted with the participants to assess their status and the timing of contracting SARS-CoV-2 after inhalation and evaluate the persistent symptoms of post-COVID-19 using a scale. The scale comprised 49 items and was used to assess the severity of the post-COVID-19 impact using 8 indicators: fatigue, shortness of breath, brain fog, impaired coordination, physical pain, impaired sleep quality, depression, and impaired quality of life. These indicators were selected based on the common symptoms of post-COVID-19 condition (PCC), that is, a set of signs and symptoms that emerge during or after an infection consistent with COVID-19 and are not explained by an alternative diagnosis [20].

Each item was rated as “never occurred,” “slightly affected,” “moderately affected,” and “severely affected,” with scores of 0, 1, 2, and 3, respectively. The Cronbach α values of the 8 indicators ranged from .79 to .94, indicating acceptable reliability [21]. The detailed questionnaire and Cronbach α values are provided in [Multimedia Appendix 1](#).

Statistical Analyses

The baseline characteristics are presented as means (SDs) for continuous variables and percentages for categorical variables. Missing values were treated and reported in all analyses.

The attenuation curves of the antibody and time were fitted through locally weighted scatterplot smoothing (LOWESS), a nonparametric method used in the analysis of local regression. The sample was divided into short intervals, and weighted polynomial fitting to the sample in each interval was conducted. Linear regressions were constructed to fit the curve stratified according to the preceding immune program.

Comparisons of geometric mean titers (GMTs) and geometric mean increases (GMIs) between the groups were performed using logarithmic conversion values. The differences in

antibodies before and after administration were compared using a paired Student 2-tailed *t* test, and linear regression was used to compare GMIs among different groups. For positive seroconversion, the antibodies after administration should increase by 4-fold or more, according to the literature [19,22]. A logistic regression was used to compare the seroconversion rate. On the basis of the marginal forecast rates of each category estimated by the regression, we used the weighted average of the standard population to calculate the direct standardized seroconversion rate and GMI. All test criteria to confirm the hypothesis were bilateral, with a significance level of .05. The adjusted α was reduced to .017 when pairwise comparisons were made between the 3 groups.

Kaplan-Meier analysis was used to plot the uninfected curves and cumulative hazard curves of different antibody levels, and the log-rank test was used to compare the difference in infection time among individuals with different levels. A multivariate Cox proportional hazard regression model was used to adjust for the effects of confounding factors on the results. Finally, a linear regression model was constructed to analyze the correlation between different sequelae scores and antibodies.

All statistical analyses were performed using R (version 4.2.0; R Development Core Team) and Stata (version 17.0; Stata Corporation).

Ethics Approval

The protocol was approved by the institutional review board of the Guizhou Center for Disease Control and Prevention (approval number Q2023-03) and was performed in accordance with the Declaration of Helsinki and the Good Clinical Practice guidelines. All participants provided written informed consent before enrollment.

Results

Basic Characteristics of the Participants

A total of 327 participants who completed the vaccination and blood sampling were enrolled in the final statistical analysis. The specific participant entry and exit processes are outlined in [Multimedia Appendix 1](#). A total of 234 female and 93 male participants were enrolled, with a mean age of 39.4 (SD 9.5) years. All participants received 3 doses of the COVID-19 vaccine before the survey, including 2 doses for basic immunization and 1 dose for booster immunization. Of the 327 participants, 166 received the BBIBP-CorV inactivated vaccine (SinoBio Pharmaceutical Ltd) and 161 received the CoronaVac inactivated vaccine produced by Sinovac Ltd. Among them, 233 participants worked in hospitals, 84 worked in nonhospital institutions, and the remaining 10 were reluctant to report their occupations and workplaces. Detailed information is provided in [Table 1](#).

Table 1. The characteristics of participants (N=327).

| Characteristics | Participants |
|-------------------------------------|--------------|
| Age (years), mean (SD) | 39.4 (9.5) |
| Age (years), n (%) | |
| <30 | 63 (19.3) |
| 30-40 | 94 (28.7) |
| 40-50 | 118 (36.1) |
| ≥50 | 52 (15.9) |
| Sex, n (%) | |
| Male | 93 (28.4) |
| Female | 234 (71.6) |
| Ethnicity, n (%) | |
| Han | 279 (85.3) |
| Minority | 48 (14.5) |
| Workplace^a, n (%) | |
| Hospital | 233 (73.5) |
| Nonhospital | 84 (26.5) |
| Primary booster vaccine | |
| BBIBP-CorV | 166 (50.8) |
| CoronaVac | 161 (49.2) |

^an=317; 10 subjects were not willing to report their occupation or workplace.

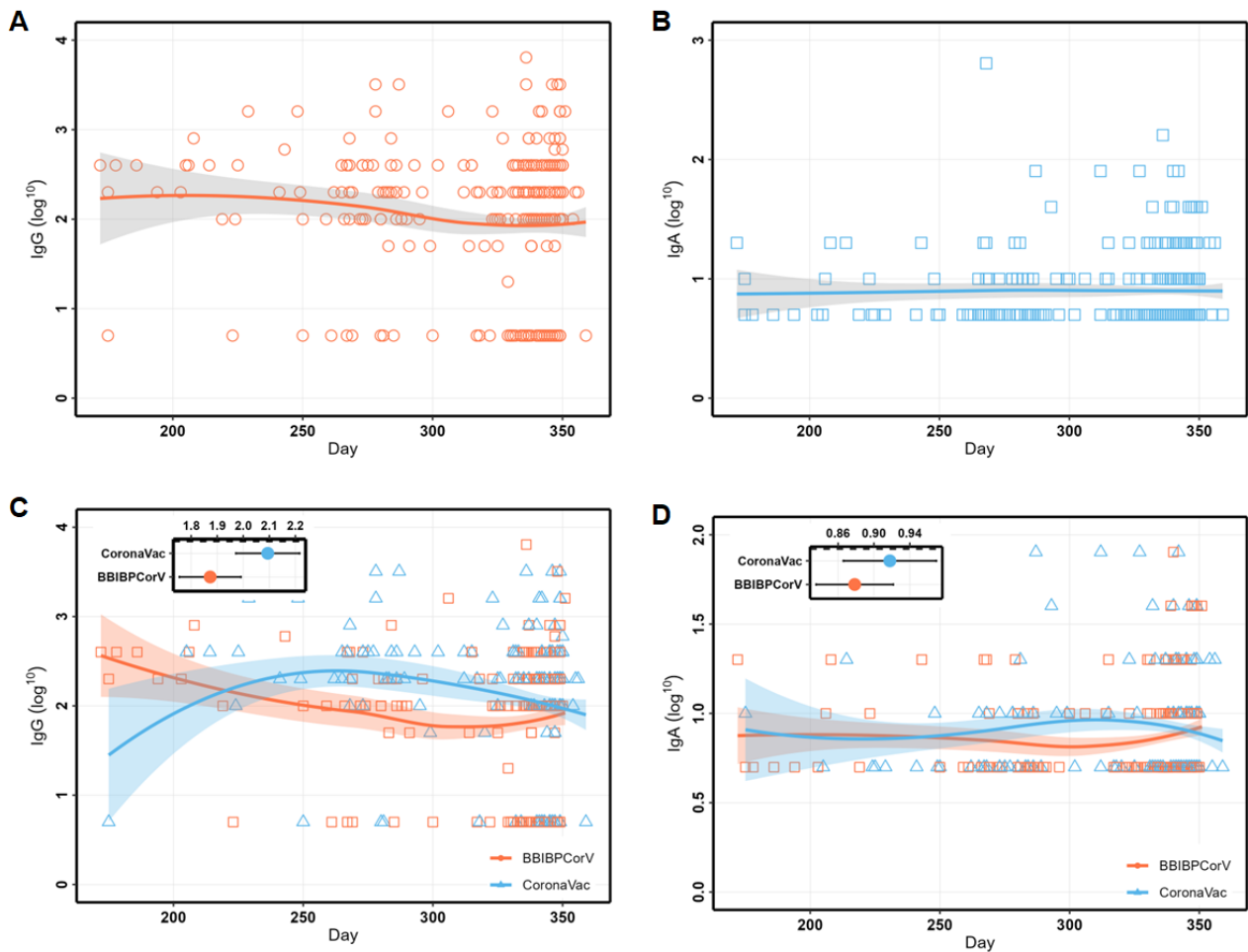
Long-Term Durability of the Antibodies After Primary-Booster Administration (Third Dose)

Before receiving the second booster (fourth dose), we measured the antibody levels of the participants. Approximately 6-12 months after receiving the booster, the GMT of the IgG antibody was 95.9 (95% CI 78.8-116.7), while that of the IgA antibody was 7.9 (95% CI 7.3-8.5). The time interval between the last booster immunization and antibody detection was used as the independent variable, while the antibody level was used as the dependent variable for LOWESS segmented curve fitting. The fitting results showed that the IgG antibody level slowly declined

with time after the primary booster, while the IgA antibody level remained almost constant (GMT 7.9; [Figure 1A](#) and [B](#)).

We proceeded to stratify the results based on the different immunization programs and perform linear fitting. [Figure 1C](#) shows that the declining trend for the IgG antibody level of participants administered the CoronaVac vaccine and those administered the BBIBP-CorV vaccine was consistent after 6 months of immunization, with the IgG antibody titer of CoronaVac (GMT 124.2; 95% CI 93.8-164.3) vaccine recipients being slightly higher than that of the BBIBP-CorV vaccine recipients (GMT 74.7; 95% CI 56.9-97.9). [Figure 1D](#) shows that the levels and trends of the 2 are almost identical.

Figure 1. The locally weighted scatterplot smoothing (LOWESS) curves of IgG and IgA over time during 6-12 months after primary booster administration. (A) The curves of IgG over time. (B) The curves of IgA over time. (C) The changing curves of IgG stratified by a booster vaccine (third dose). (D) The changing curves of IgA stratified by a booster vaccine (third dose). IgA: immunoglobulin A; IgG: immunoglobulin G.



Immunogenicity of the Second Booster (Fourth Dose)

At 28 days after receiving the inhalant vaccine as the second booster, the GMT of the IgG antibody was 5066.5 (95% CI 4418.1-5810.1), while that of the IgA antibody was 108.6 (95% CI 95.4-123.5). The GMI of IgG was 52.8 (95% CI 42.6-65.6), and the seroconversion rate reached 94.5% (95% CI 92-97). The GMI of IgA was 13.7 (95% CI 12-15.7), and the antibody seroconversion rate was 89.3% (95% CI 85.9-92.7).

The antibody levels of individuals with different sociodemographic characteristics and prevaccination programs showed significant improvement after vaccination. The line

graphs of the pre and postvaccination GMTs for the IgG and IgA antibodies of different groups are shown in [Figure 2](#).

We explored the crude and adjusted changes in antibody (GMI and seroconversion rate) among different groups based on age, sex, ethnicity, workplace, and type of primary booster vaccine, as detailed in [Table 2](#). There was no difference in postvaccination GMI or seroconversion of IgG antibodies among the different demographic groups. However, for participants who received BBIBP-CorV (adjusted rate 97.4%; 95% CI 94.9-99.9) as their initial booster, the seroconversion rate of the IgG antibody was higher than that of those who received CoronaVac (adjusted rate 91.4%; 95% CI 87.2-95.7) after the secondary booster immunization.

Figure 2. The differences of the pre- and postvaccination GMTs for the IgG and IgA antibodies of different groups. (A-E) GMTs for the IgG. (A) Group by age. (B) Group by sex. (C) Group by ethnicity. (D) Group by type of primary booster vaccine. (E) Group by workplace. (F-J) GMTs for the IgA. (F) Group by age. (G) Group by sex. (H) Group by ethnicity. (I) Group by type of primary booster vaccine. (J) Group by workplace. GMT: geometric mean titer; IgA: immunoglobulin A; IgG: immunoglobulin G.

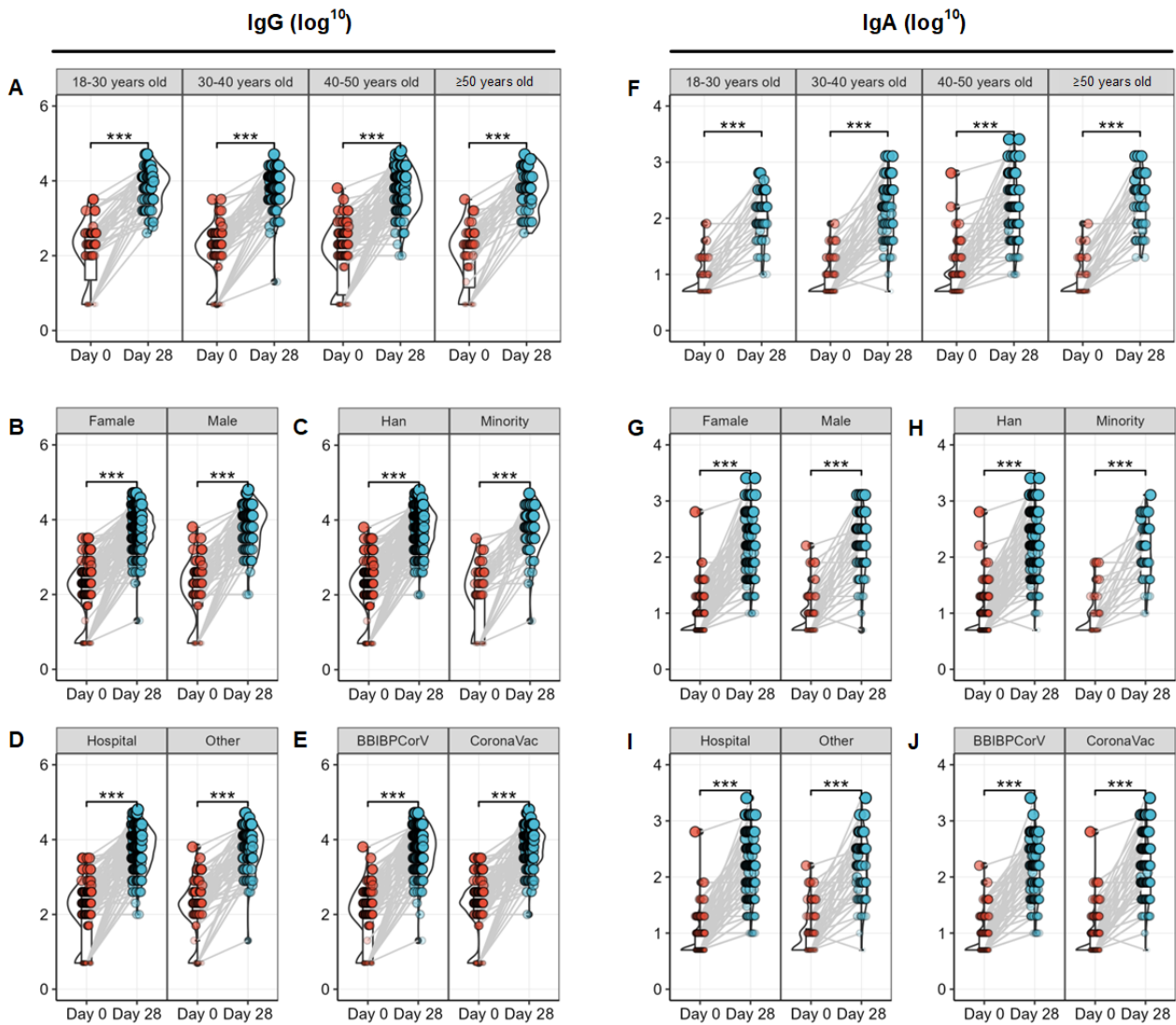


Table 2. The distribution and difference of geometric mean increase and positive seroconversion rate of immunoglobulin G and immunoglobulin A antibodies among groups.

| Variable | Geometric mean increase, mean (95% CI) | | | Seroconversion rate, proportion (95% CI) | | |
|-------------------------|--|-------------------|----------------|--|-------------------|----------------|
| | Crude | Adjusted | <i>P</i> value | Crude | Adjusted | <i>P</i> value |
| Immunoglobulin G | | | | | | |
| Overall | 52.8 (42.6-65.6) | N/A ^a | N/A | 94.5 (92-97) | N/A | N/A |
| Age (years) | | | | | | |
| <30 | 56.6 (33.9-94.7) | 53.6 (32.2-89.1) | — ^b | 93.7 (87.6-99.7) | 92.3 (85-99.7) | — |
| 30-40 | 49.1 (35.1-68.6) | 52.2 (34.4-79.3) | .94 | 97.9 (94.9-100.8) | 97.7 (94.5-100.9) | .16 |
| 40-50 | 48.4 (32.2-72.6) | 49.9 (34.3-72.5) | .83 | 91.5 (86.5-96.6) | 92.3 (87.6-96.9) | .99 |
| ≥50 | 67.7 (41-111.9) | 69.8 (39.8-122.7) | .49 | 96.2 (90.9-101.5) | 95.9 (90.5-101.4) | .45 |
| Sex | | | | | | |
| Male | 52 (34.6-78.1) | 57.6 (37.4-88.8) | — | 92.5 (87.1-97.9) | 93.3 (88.3-98.2) | — |
| Female | 53.2 (41.2-68.6) | 52.6 (40.3-68.6) | .73 | 95.3 (92.6-98) | 94.8 (91.8-97.8) | .59 |
| Ethnicity | | | | | | |
| Han | 52.8 (41.9-66.5) | 53.5 (42.1-68) | — | 94.3 (91.5-97) | 94.2 (91.5-96.9) | — |
| Minority | 52.9 (28.6-97.9) | 56.8 (31.6-102.1) | .85 | 95.8 (90.1-101.6) | 95 (88.3-101.7) | .84 |
| Workplace | | | | | | |
| Hospital | 61.3 (47-80.1) | 61.5 (47.4-79.8) | — | 94.4 (91.5-97.4) | 94.4 (91.5-97.3) | — |
| Nonhospital | 37.9 (26-55.1) | 37.6 (24.2-58.4) | .06 | 94 (88.9-99.2) | 94.1 (89-99.2) | .92 |
| Booster vaccine | | | | | | |
| BBIBP-CorV | 64.1 (48-85.6) | 65.7 (48-89.9) | — | 97.6 (95.2-99.9) | 97.4 (94.9-99.9) | — |
| CoronaVac | 43.3 (31.4-59.6) | 44.3 (32.4-60.7) | .09 | 91.3 (86.9-95.7) | 91.4 (87.2-95.7) | .03 |
| Immunoglobulin A | | | | | | |
| Overall | 13.7 (12-15.7) | N/A | N/A | 89.3 (85.9-92.7) | N/A | N/A |
| Age (years) | | | | | | |
| <30 | 12.9 (9.8-16.8) | 13.2 (9.7-18) | — | 88.9 (81-96.7) | 89.2 (81.5-96.9) | — |
| 30-40 | 14.1 (11.2-17.7) | 14.8 (11.5-19.1) | .58 | 92.6 (87.2-97.9) | 93.7 (88.8-98.6) | .31 |
| 40-50 | 11.8 (9.3-14.9) | 11.1 (8.9-14) | .38 | 83.9 (77.2-90.6) | 82.5 (75.3-89.8) | .25 |
| ≥50 | 20.3 (14.6-28.2) | 20.4 (14.5-28.7) | .07 | 96.2 (90.9-101.5) | 95.9 (90.3-101.5) | .21 |
| Sex | | | | | | |
| Male | 16.5 (12.6-21.5) | 16.9 (13-21.9) | — | 89.2 (82.9-95.6) | 89.7 (83.4-95.9) | — |
| Female | 12.8 (11-14.8) | 12.6 (10.8-14.9) | .07 | 89.3 (85.3-93.3) | 89.1 (85-93.2) | .88 |
| Ethnicity | | | | | | |
| Han | 13.9 (12.1-16) | 13.8 (11.9-15.9) | — | 90.3 (86.8-93.8) | 90.1 (86.6-93.6) | — |
| Minority | 12.8 (8.6-19.2) | 13.6 (9.5-19.3) | .94 | 83.3 (72.6-94) | 83.9 (73.2-94.7) | .22 |
| Workplace | | | | | | |
| Hospital | 14 (12-16.4) | 14.3 (12.2-16.7) | — | 89.7 (85.8-93.6) | 89.8 (86-93.6) | — |
| Nonhospital | 13 (9.8-17.1) | 12.3 (9.4-16.1) | .36 | 88.1 (81.1-95.1) | 87.8 (80.6-94.9) | .62 |
| Booster vaccine | | | | | | |
| BBIBP-CorV | 13.3 (11-15.9) | 13.1 (10.8-15.8) | — | 88.6 (83.7-93.4) | 87.6 (82.5-92.8) | — |
| CoronaVac | 14.2 (11.8-17.2) | 14.4 (11.9-17.5) | .48 | 90.1 (85.4-94.7) | 90.8 (86.4-95.2) | .36 |

^aN/A: not applicable.^b—: not available.

Real-World Protective Effect of Secondary Booster Administration Against Infection

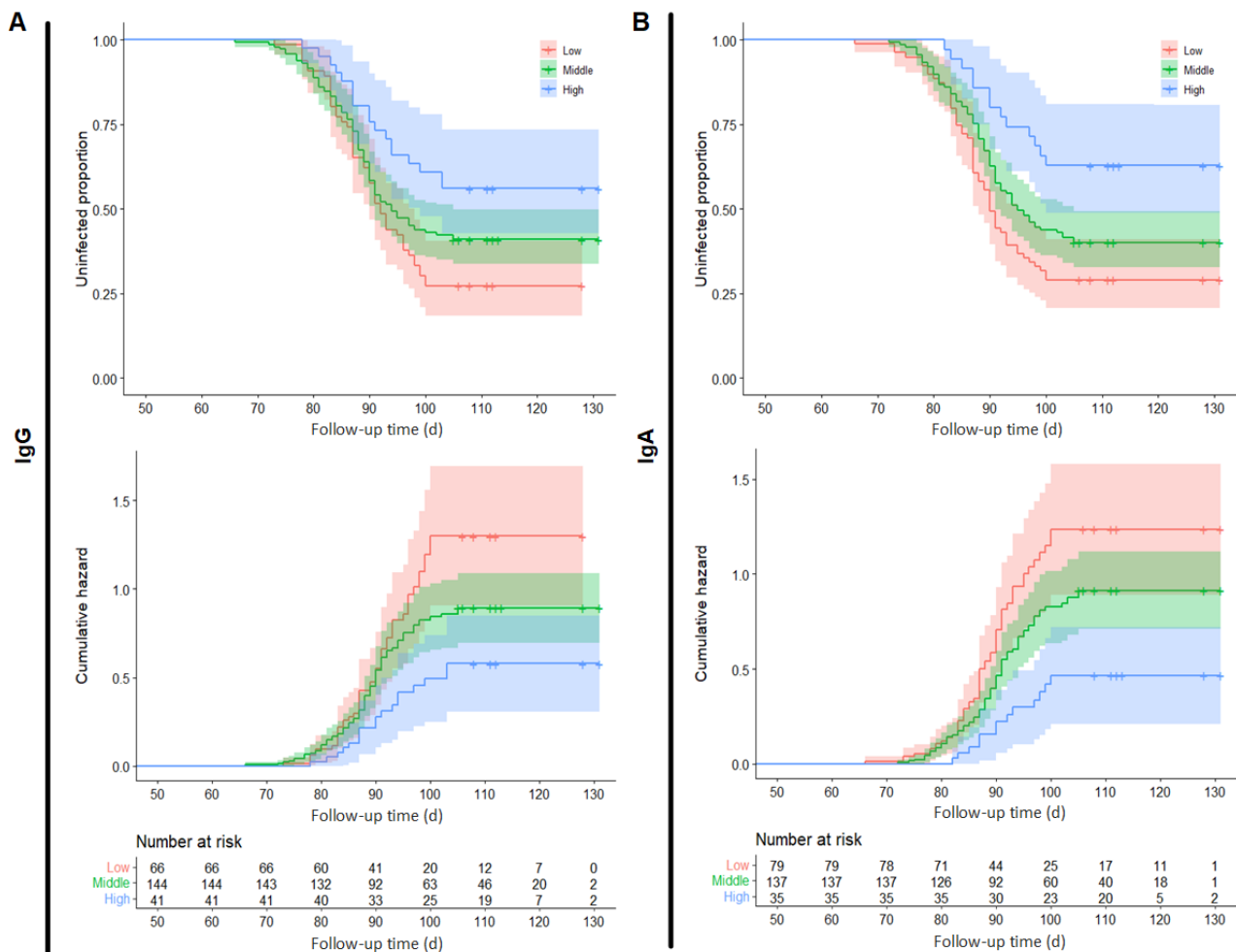
After the second booster, we continued to track and follow up with the participants to derive the breakthrough infection rates. As of January 30, 2022, we followed up with 251 participants; the breakthrough infection rate was 60.2% (151/251).

After adjusting for confounding factors, Cox regression analysis revealed a significant correlation between postimmunization IgG antibody levels and breakthrough COVID-19 infection (hazard ratio 0.60; 95% CI 0.45-0.79); the same result was obtained for IgA antibody levels (hazard ratio 0.55; 95% CI 0.39-0.78). This result suggests that both IgG and IgA antibodies after the second booster can provide a certain degree of protection against COVID-19 and prevent infection. We categorized postimmunization IgG and IgA antibodies into 3 levels, high, medium, and low, based on 25th percentile and

75th percentile. Survival and cumulative risk curves were plotted for each level, as shown in Figure 3. Higher levels of postimmunization IgG antibodies were associated with better protection against COVID-19 infection and a lower cumulative risk for vaccine recipients (Figure 3A). Similarly, higher levels of postimmunization IgA antibodies were associated with better protection against COVID-19 infection and a lower cumulative risk for vaccine recipients (Figure 3B).

Kaplan-Meier curves for the risk of COVID-19 infection for age and sex subgroups were plotted, as shown in Figure 1. Differences between age groups were found: participants between 30 and 40 years of age had a higher hazard risk than other age groups. The differences were insignificant after adjusting for other sociodemographic factors. Women appear to have a higher risk of infection than men, but this did not reach statistical significance ($P=.051$).

Figure 3. The Kaplan-Meier and cumulative hazard curve of COVID-19 infection grouped by antibody levels. (A) Curves grouped by IgG levels (low, middle, and high). (B) Curves grouped by IgA levels (low, middle, and high). IgA: immunoglobulin A; IgG: immunoglobulin G.



Real-World Protective Effect of Secondary Booster Administration Against Post-COVID-19 Symptoms

Finally, a questionnaire was used to evaluate the post-COVID-19 symptoms of 151 participants who experienced breakthrough infections, including fatigue, shortness of breath, brain fog, impaired coordination, physical pain, impaired sleep quality, depression, and impaired quality of life. The most

common symptoms were fatigue, impaired sleep quality, and impaired quality of life, with mean scores of 0.46, 0.24, and 0.20, respectively.

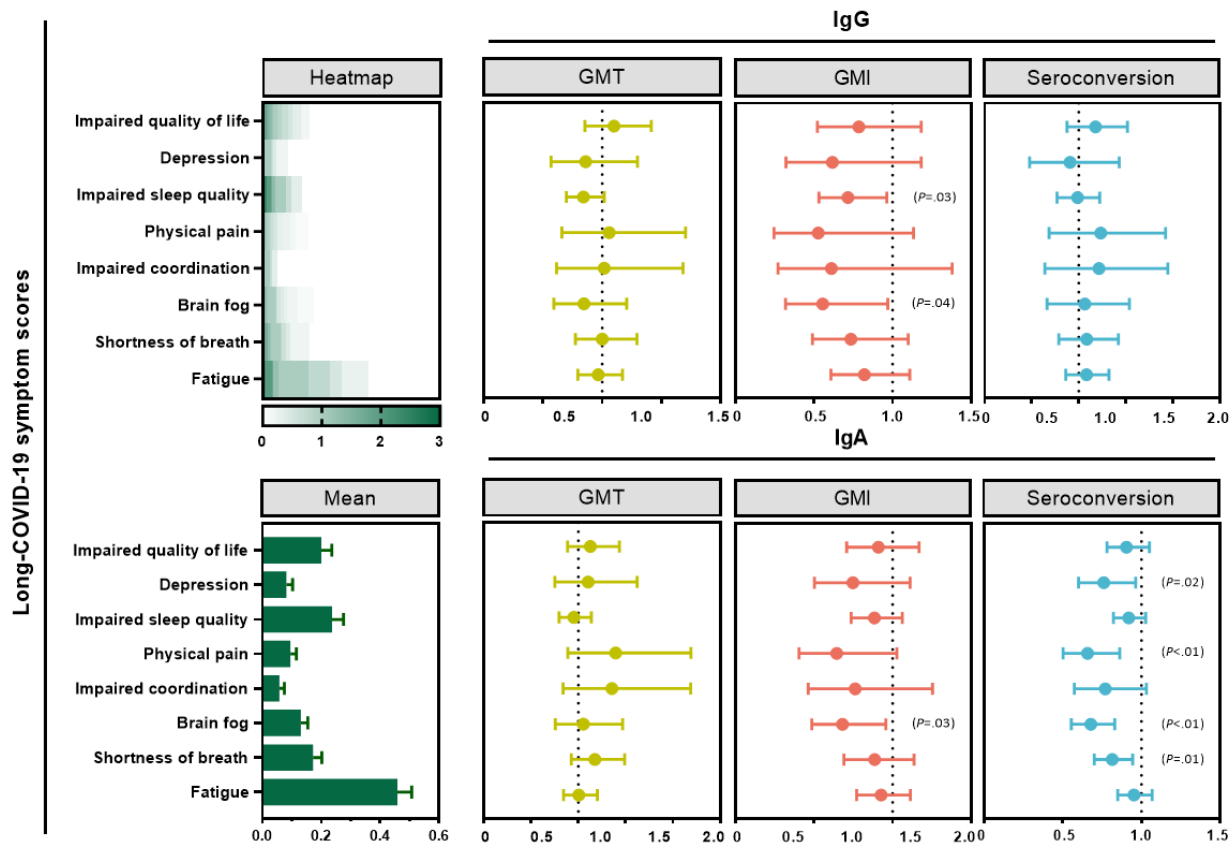
We constructed a regression model with postimmunization antibody levels as the independent variable and questionnaire scores as the dependent variable. Based on the results, the GMI levels of postimmunization IgG antibodies had a protective effect on brain fog (odds ratio [OR] 0.56; 95% CI 0.32-0.97)

and sleep quality (OR 0.71; 95% CI 0.53-0.96). The GMI levels of postimmunization IgA antibodies had a protective effect on brain fog (OR 0.68; 95% CI 0.49-0.96). Additionally, participants who were IgA antibody seropositive after infection had milder symptoms of shortness of breath (OR 0.81; 95% CI 0.7-0.95), brain fog (OR 0.68; 95% CI 0.55-0.83), impaired

coordination (OR 0.66; 95% CI 0.50-0.86), and physical pain (OR 0.76; 95% CI 0.60-0.96; Figure 4).

The age-stratified protective effect of IgG and IgA after secondary enhancement on post-COVID-19 symptoms was reported. Unfortunately, no differences between age groups were found, and detailed results can be seen in Figure 2.

Figure 4. The adjusted regression models of antibodies on post-COVID-19 scores. The heatmap and bar chart on the left represent the discrete and central trends of the post-COVID-19 scores, and the scores of each subitem entered the model as a dependent variable. The 3 graphs on the top right are the results of entering different metrics of IgG (ie, GMT, GMI, and seroconversion rate) into the regression model as independent variables. The 3 graphs on the bottom right are the results of entering different metrics of IgA (ie, GMT, GMI, and seroconversion rate) into the regression model as independent variables. Each regression adjusted for basic sociodemographic characteristics. GMI: geometric mean increase; GMT: geometric mean titer; IgA: immunoglobulin A; IgG: immunoglobulin G.



Discussion

Overview

Based on existing evidence, there is a significant decline in neutralizing antibody titers 4-5 months after completion of the routine vaccination program [23,24]. The immune protection induced by the vaccine declines continuously with time, highlighting the urgent need for booster vaccination to enhance protection. According to statistics from the National Health Commission, approximately 850 million people in China have received their booster vaccine as of February 2023 [16]. The administration of a third dose of the same vaccine has been demonstrated to significantly increase neutralizing antibody levels and effectively reduce the symptomatic infection rate of the SARS-CoV-2 variant [25,26].

The literature suggests a decline or even disappearance of antibody levels within a short period of 3-6 months [27-30]. In

this study, we tracked the antibody level owing to the first booster for 6-12 months without the interference of natural infection to evaluate long-term immunogenicity. We found that the decline in antibody level was slow after 6 months and maintained at a relatively low level, with a GMT of approximately 96 for IgG and approximately 8 for IgA. In confirmatory research, vaccine effectiveness was estimated to decline from approximately 70% one week after the booster dose to approximately 40% at 15 weeks or more [31].

The titers of postadministration antibodies vary according to the vaccine type. In China, inactivated vaccines are the most commonly used vaccines for basic and booster immunizations owing to their safety and stability. Based on evidence from the Chinese Center for Disease Control and Prevention, the GMTs owing to BBIBP-CorV were 25 at 1 month and 4 at 12 months, while those owing to CoronaVac were 20.2 and 4.1, respectively [32]. In this study, ELISA revealed that the IgG antibody titers

from CoronaVac were slightly higher than those from BBIBP-CorV at 6-12 months after administration but converged at 12 months.

Breakthrough infections have become more common with the decline in antibody levels and the development of new VOCs with strong immunologic escape, despite the remarkable effect of primary boosting [33]. A fourth dose of the COVID-19 vaccine can boost cellular and humoral immunity, and the peak responses were found to be similar to the peak responses after the third dose [34]. According to some clinical trials, the adenovirus vector booster dose based on an inactivated vaccine could lead to higher neutralization antibodies than homologous boosting [35,36]. In China, nearly 47 million residents have now completed the sequential booster immunization since the start of its dissemination in November 2022 [14].

Based on the available evidence, a prospective study was conducted to evaluate heterologous secondary booster administration. To our knowledge, this study is the first real-world evaluation of the effectiveness of the second booster dose in China. Herein, the inhalant Ad5-nCoV vaccine was administered as a second booster dose. Inhalant Ad5-nCoV is homologous to injectable Ad5-nCoV but achieves protection through mucosal immunization through inhalation. Mucosal immunity is a critical component of the human immune system, with more than 90% of infections occurring in the mucosa, which comprises numerous dendritic cells with strong T-cell activation capacity that can induce an immune response. ELISA to detect the serum antibodies revealed that the GMT of the IgG antibody was 4978.2 and that of the IgA antibody was 107.8 at 28 days after inhalation. The seroconversion rates for IgG and IgA were 93.8% and 86.9%, respectively. The antibody titers of the primary booster against the Omicron variant were attenuated relative to those of other virus strains, such as the wild type and other VOCs. Therefore, the use of heterogeneous vaccines for the second-booster procedures is a concern for the prevention of the Omicron variant [37].

China suffered a COVID-19 epidemic between December 2022 and January 2023 owing to changes in health policies and the impact of the Omicron variant, with a peak in cases on December 22, 2022 [13]. In this study, we followed up with participants for 4 months after receiving their fourth vaccine dose to assess their real-world COVID-19 infection status. As of January 30, 2023, 60.2% (151/251) of participants self-reported that they had been infected with COVID-19. Notably, IgG and IgA provided strong protection against

COVID-19 infection, as demonstrated by the high antibody titers after the fourth dose.

In addition to infection prevention, the long-term effects of COVID-19 are also concerning. Studies from high-income countries suggest that vaccination may alleviate post-COVID-19 or PCC [38]. However, based on other evidence, COVID-19 vaccination is not associated with improvement in PCC [39,40]. We sought to assess some nonspecific post-COVID-19 symptoms. Based on our results, the GMI and seroconversion rates of IgG and IgA may alleviate some of the symptoms after secondary booster administration, including sleep quality, shortness of breath, brain fog, impaired coordination, and physical pain. The GMTs had no statistical relationship, which may be explained by the insufficient statistical efficacy induced by the small sample size. However, its clinical value is still worth exploring. The varying levels of IgG and IgA suggest the mechanisms of immune protection after infection, suggesting that further tracking and research are warranted.

Our study had some limitations. First, the SARS-CoV-2 infection of participants was self-reported, which may have led to recall and reporting biases. Further, asymptomatic patients may not have been identified. Second, the antibody detection method used was an ELISA quantitative assay rather than the neutralization test. However, the results of both tests are highly correlated according to the literature. ELISA can be used as a substitute for the gold standard to assess immunogenicity [22]. Third, there was no control group setting without a second booster dose for estimating the vaccine effect of the second booster dose. Their use of antibody levels was reasonable in accordance with previous trials, however [41]. Finally, only follow-up data collected within approximately 1 month after infection were reported in this study. These findings may offer suggestions for future populations with post-COVID-19. Further studies are needed, including appropriate control measures for unvaccinated individuals, to confirm the trajectory of persistent symptoms after COVID-19 vaccination.

Conclusions

The IgG and IgA antibodies did not decrease significantly but remained at a relatively low level after administration of the second booster dose. The antibodies generated significant immunogenic protection against breakthrough infections and might partially alleviate post-COVID-19 symptoms. Further studies on secondary booster administration are needed to validate these correlations.

Acknowledgments

The authors would like to thank professors Fengcai Zhu and Aihua Zhang (Guizhou Medical University) for their great support and expertise in this study. We would also like to thank the following organizations for their full support in the implementation of the project: Guiyang Bureau of Health, Jinhua Community Healthcare Center of Guanshanhu District, Century-City Third Community Healthcare Center of Guanshanhu District, Qingzhen Center for Disease Control and Prevention, Baihua Community Healthcare Center of Qingzhen, Hongta Community Healthcare Center of Qingzhen, Baiyun Branch of Guizhou Provincial People's Hospital, the Affiliated Hospital of Guizhou Medical University (Baiyun Branch), Kaiyang Center for Disease Control and Prevention, Kaiyang COVID-19 Vaccination Site, Xifeng Bureau of Health, Xifeng Maternity and Child Healthcare Hospital, Xiuwen Center for Disease Control and Prevention, Baixin Hospital of Xiuwen, Center for Disease Control and Prevention of Yunyan District, and Qianling Hospital of Yunyan District. This study was supported by the National Key Research and Development Program of China (2021YFC2301604) and Joint Research Fund for Beijing Natural Science Foundation and Haidian

Original Innovation (L222029). The funders had no role in study design, data collection and analysis, the decision to publish, or the preparation of the manuscript.

Data Availability

All data sets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

GY, FC, and SL designed the study. ML designed the survey instrument. QM, RZ, CL, FX, LL, LZ, SZ, and NH all contributed to the conceptual design of the research and to the data collection. XC and TZ conducted the statistical analysis. ML and TZ wrote the manuscript. GY (ghyang_gzmu@outlook.com), FC (cuifuq@bjmu.edu.cn), and SL (ShiguangLei193@foxmail.com) are co-corresponding authors for this article.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Post-COVID symptom scale, Kaplan-Meier curves of COVID-19 infection, and age-stratified correlation between post-administrated antibody and post-COVID symptoms.

[\[DOCX File , 7592 KB-Multimedia Appendix 1\]](#)

References

1. Haldane V, De Foo C, Abdalla SM, Jung A, Tan M, Wu S, et al. Health systems resilience in managing the COVID-19 pandemic: lessons from 28 countries. *Nat Med* 2021;27(6):964-980 [FREE Full text] [doi: [10.1038/s41591-021-01381-y](https://doi.org/10.1038/s41591-021-01381-y)] [Medline: [34002090](https://pubmed.ncbi.nlm.nih.gov/34002090/)]
2. Tracking SARS-CoV-2 variants. World Health Organization. 2022. URL: <https://www.who.int/activities/tracking-SARS-CoV-2-variants> [accessed 2022-02-24]
3. Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, et al. Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* 2021;592(7854):438-443 [FREE Full text] [doi: [10.1038/s41586-021-03402-9](https://doi.org/10.1038/s41586-021-03402-9)] [Medline: [33690265](https://pubmed.ncbi.nlm.nih.gov/33690265/)]
4. Faria NR, Mellan TA, Whittaker C, Claro IM, Candido DDS, Mishra S, et al. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science* 2021;372(6544):815-821 [FREE Full text] [doi: [10.1126/science.abb2644](https://doi.org/10.1126/science.abb2644)] [Medline: [33853970](https://pubmed.ncbi.nlm.nih.gov/33853970/)]
5. Dhar MS, Marwal R, Vs R, Ponnusamy K, Jolly B, Bhojar RC, Indian SARS-CoV-2 Genomics Consortium (INSACOG), et al. Genomic characterization and epidemiology of an emerging SARS-CoV-2 variant in Delhi, India. *Science* 2021;374(6570):995-999 [FREE Full text] [doi: [10.1126/science.abj9932](https://doi.org/10.1126/science.abj9932)] [Medline: [34648303](https://pubmed.ncbi.nlm.nih.gov/34648303/)]
6. Viana R, Moyo S, Amoako DG, Tegally H, Scheepers C, Althaus CL, et al. Rapid epidemic expansion of the SARS-CoV-2 omicron variant in Southern Africa. *Nature* 2022;603(7902):679-686 [FREE Full text] [doi: [10.1038/s41586-022-04411-y](https://doi.org/10.1038/s41586-022-04411-y)] [Medline: [35042229](https://pubmed.ncbi.nlm.nih.gov/35042229/)]
7. Cele S, Jackson L, Khoury DS, Khan K, Moyo-Gwete T, Tegally H, NGS-SA, COMMIT-KZN Team, et al. Omicron extensively but incompletely escapes Pfizer BNT162b2 neutralization. *Nature* 2022;602(7898):654-656 [FREE Full text] [doi: [10.1038/s41586-021-04387-1](https://doi.org/10.1038/s41586-021-04387-1)] [Medline: [35016196](https://pubmed.ncbi.nlm.nih.gov/35016196/)]
8. Zhang L, Li Q, Liang Z, Li T, Liu S, Cui Q, et al. The significant immune escape of pseudotyped SARS-CoV-2 variant Omicron. *Emerg Microbes Infect* 2022;11(1):1-5 [FREE Full text] [doi: [10.1080/22221751.2021.2017757](https://doi.org/10.1080/22221751.2021.2017757)] [Medline: [34890524](https://pubmed.ncbi.nlm.nih.gov/34890524/)]
9. Lu L, Mok BWY, Chen LL, Chan JMC, Tsang OTY, Lam BHS, et al. Neutralization of severe acute respiratory syndrome coronavirus 2 omicron variant by sera from BNT162b2 or CoronaVac vaccine recipients. *Clin Infect Dis* 2022;75(1):e822-e826 [FREE Full text] [doi: [10.1093/cid/ciab1041](https://doi.org/10.1093/cid/ciab1041)] [Medline: [34915551](https://pubmed.ncbi.nlm.nih.gov/34915551/)]
10. Dejnirattisai W, Shaw RH, Supasa P, Liu C, Stuart AS, Pollard AJ, Com-COV2 study group. Reduced neutralisation of SARS-CoV-2 omicron B.1.1.529 variant by post-immunisation serum. *Lancet* 2022;399(10321):234-236 [FREE Full text] [doi: [10.1016/S0140-6736\(21\)02844-0](https://doi.org/10.1016/S0140-6736(21)02844-0)] [Medline: [34942101](https://pubmed.ncbi.nlm.nih.gov/34942101/)]
11. Edara VV, Manning KE, Ellis M, Lai L, Moore KM, Foster SL, et al. mRNA-1273 and BNT162b2 mRNA vaccines have reduced neutralizing activity against the SARS-CoV-2 omicron variant. *Cell Rep Med* 2022;3(2):100529 [FREE Full text] [doi: [10.1016/j.xcrm.2022.100529](https://doi.org/10.1016/j.xcrm.2022.100529)] [Medline: [35233550](https://pubmed.ncbi.nlm.nih.gov/35233550/)]
12. Ai J, Zhang H, Zhang Y, Lin K, Zhang Y, Wu J, et al. Omicron variant showed lower neutralizing sensitivity than other SARS-CoV-2 variants to immune sera elicited by vaccines after boost. *Emerg Microbes Infect* 2022;11(1):337-343 [FREE Full text] [doi: [10.1080/22221751.2021.2022440](https://doi.org/10.1080/22221751.2021.2022440)] [Medline: [34935594](https://pubmed.ncbi.nlm.nih.gov/34935594/)]

13. The situation of the SARS-CoV-2 infection in China. Chinese Center for Disease Control and Prevention. 2023. URL: https://www.chinacdc.cn/jkzt/crb/zl/szkb_11803/jszl_13141/202302/t20230218_263807.html [accessed 2023-02-24]
14. Dedicated COVID-19 vaccination dashboard. World Health Organization. 2023. URL: <https://covid19.who.int> [accessed 2023-02-24]
15. Bai Y, Peng Z, Wei F, Jin Z, Wang J, Xu X, et al. Study on the COVID-19 epidemic in mainland China between November 2022 and January 2023, with prediction of its tendency. *J Biosaf Biosecur* 2023;5(1):39-44 [FREE Full text] [doi: [10.1016/j.jobb.2023.03.001](https://doi.org/10.1016/j.jobb.2023.03.001)] [Medline: [36992708](https://pubmed.ncbi.nlm.nih.gov/36992708/)]
16. Transcript of press conference under the Joint Prevention and Control Mechanism of The State Council on 23 Feb 2023. National Health Commission of China. URL: <http://www.nhc.gov.cn/xcs/yqfkdt/202302/172708cde8fb4e40976e693443bbb596.shtml> [accessed 2023-02-24]
17. WHO interim recommendations for heterologous COVID-19 vaccine schedules. World Health Organization. 2021. URL: <https://www.who.int/publications/i/item/WHO-2019-nCoV-vaccines-SAGE-recommendation-heterologous-schedules> [accessed 2023-08-26]
18. Jin P, Guo X, Chen W, Ma S, Pan H, Dai L, et al. Safety and immunogenicity of heterologous boost immunization with an adenovirus type-5-vectored and protein-subunit-based COVID-19 vaccine (Convidecia/ZF2001): a randomized, observer-blinded, placebo-controlled trial. *PLoS Med* 2022;19(5):e1003953 [FREE Full text] [doi: [10.1371/journal.pmed.1003953](https://doi.org/10.1371/journal.pmed.1003953)] [Medline: [35617368](https://pubmed.ncbi.nlm.nih.gov/35617368/)]
19. Wu S, Huang J, Zhang Z, Wu J, Zhang J, Hu H, et al. Safety, tolerability, and immunogenicity of an aerosolised adenovirus type-5 vector-based COVID-19 vaccine (Ad5-nCoV) in adults: preliminary report of an open-label and randomised phase 1 clinical trial. *Lancet Infect Dis* 2021;21(12):1654-1664 [FREE Full text] [doi: [10.1016/S1473-3099\(21\)00396-0](https://doi.org/10.1016/S1473-3099(21)00396-0)] [Medline: [34324836](https://pubmed.ncbi.nlm.nih.gov/34324836/)]
20. Fawzy NA, Shaar BA, Taha RM, Arabi TZ, Sabbah BN, Alkodaymi MS, et al. A systematic review of trials currently investigating therapeutic modalities for post-acute COVID-19 syndrome and registered on WHO International Clinical Trials Platform. *Clin Microbiol Infect* 2023;29(5):570-577 [FREE Full text] [doi: [10.1016/j.cmi.2023.01.007](https://doi.org/10.1016/j.cmi.2023.01.007)] [Medline: [36642173](https://pubmed.ncbi.nlm.nih.gov/36642173/)]
21. de Vet HCW, Terwee CB, Mokkink LB, Knol DL. *Measurement in Medicine: A Practical Guide*. Cambridge, UK: Cambridge University Press; 2011.
22. Dolscheid-Pommerich R, Bartok E, Renn M, Kümmerer BM, Schulte B, Schmithausen RM, et al. Correlation between a quantitative anti-SARS-CoV-2 IgG ELISA and neutralization activity. *J Med Virol* 2022;94(1):388-392 [FREE Full text] [doi: [10.1002/jmv.27287](https://doi.org/10.1002/jmv.27287)] [Medline: [34415572](https://pubmed.ncbi.nlm.nih.gov/34415572/)]
23. Collier ARY, Yu J, McMahan K, Liu J, Chandrashekar A, Maron JS, et al. Differential kinetics of immune responses elicited by Covid-19 vaccines. *N Engl J Med* 2021;385(21):2010-2012 [FREE Full text] [doi: [10.1056/NEJMc2115596](https://doi.org/10.1056/NEJMc2115596)] [Medline: [34648703](https://pubmed.ncbi.nlm.nih.gov/34648703/)]
24. Suo S, Hongyang Y, Qian L, Yang Y, Jiayin S, Jianqing X, et al. Expert recommendations on booster immunization strategies of 2019-nCoV vaccines. *Chin J Clin Infect Dis* 2022;15(3):176-184 [FREE Full text] [doi: [10.3760/cma.j.issn.1674-2397.2022.03.003](https://doi.org/10.3760/cma.j.issn.1674-2397.2022.03.003)]
25. Cao Y, Wang X, Li S, Dong Y, Liu Y, Li J, et al. A third high dose of inactivated COVID-19 vaccine induces higher neutralizing antibodies in humans against the Delta and Omicron variants: a randomized, double-blinded clinical trial. *Sci China Life Sci* 2022;65(8):1677-1679 [FREE Full text] [doi: [10.1007/s11427-022-2110-1](https://doi.org/10.1007/s11427-022-2110-1)] [Medline: [35441932](https://pubmed.ncbi.nlm.nih.gov/35441932/)]
26. Zeng G, Wu Q, Pan H, Li M, Yang J, Wang L, et al. Immunogenicity and safety of a third dose of CoronaVac, and immune persistence of a two-dose schedule, in healthy adults: interim results from two single-centre, double-blind, randomised, placebo-controlled phase 2 clinical trials. *Lancet Infect Dis* 2022;22(4):483-495 [FREE Full text] [doi: [10.1016/S1473-3099\(21\)00681-2](https://doi.org/10.1016/S1473-3099(21)00681-2)] [Medline: [34890537](https://pubmed.ncbi.nlm.nih.gov/34890537/)]
27. Cheng ZJ, Huang H, Zheng P, Xue M, Ma J, Zhan Z, et al. Humoral immune response of BBIBP COVID-19 vaccination before and after the booster immunization. *Allergy* 2022;77(8):2404-2414 [FREE Full text] [doi: [10.1111/all.15271](https://doi.org/10.1111/all.15271)] [Medline: [35255171](https://pubmed.ncbi.nlm.nih.gov/35255171/)]
28. Wisniewski AV, Luna JC, Redlich CA. Human IgG and IgA responses to COVID-19 mRNA vaccines. *PLoS One* 2021;16(6):e0249499 [FREE Full text] [doi: [10.1371/journal.pone.0249499](https://doi.org/10.1371/journal.pone.0249499)] [Medline: [34133415](https://pubmed.ncbi.nlm.nih.gov/34133415/)]
29. Lyke KE, Atmar RL, Islas CD, Posavad CM, Szydlo D, Chourdury RP, DMID 21-0012 Study Group. Rapid decline in vaccine-boosted neutralizing antibodies against SARS-CoV-2 omicron variant. *Cell Rep Med* 2022;3(7):100679 [FREE Full text] [doi: [10.1016/j.xcrm.2022.100679](https://doi.org/10.1016/j.xcrm.2022.100679)] [Medline: [35798000](https://pubmed.ncbi.nlm.nih.gov/35798000/)]
30. Thomas SJ, Moreira ED, Kitchin N, Absalon J, Gurtman A, Lockhart S, C4591001 Clinical Trial Group. Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine through 6 months. *N Engl J Med* 2021;385(19):1761-1773 [FREE Full text] [doi: [10.1056/NEJMoa2110345](https://doi.org/10.1056/NEJMoa2110345)] [Medline: [34525277](https://pubmed.ncbi.nlm.nih.gov/34525277/)]
31. Kirsebom FCM, Andrews N, Stowe J, Toffa S, Sachdeva R, Gallagher E, et al. COVID-19 vaccine effectiveness against the omicron (BA.2) variant in England. *Lancet Infect Dis* 2022;22(7):931-933 [FREE Full text] [doi: [10.1016/S1473-3099\(22\)00309-7](https://doi.org/10.1016/S1473-3099(22)00309-7)] [Medline: [35623379](https://pubmed.ncbi.nlm.nih.gov/35623379/)]

32. Wang F, Huang B, Lv H, Feng L, Ren W, Wang X, et al. Factors associated with neutralizing antibody levels induced by two inactivated COVID-19 vaccines for 12 months after primary series vaccination. *Front Immunol* 2022;13:967051 [FREE Full text] [doi: [10.3389/fimmu.2022.967051](https://doi.org/10.3389/fimmu.2022.967051)] [Medline: [36159863](https://pubmed.ncbi.nlm.nih.gov/36159863/)]
33. Levin EG, Lustig Y, Cohen C, Fluss R, Indenbaum V, Amit S, et al. Waning immune humoral response to BNT162b2 Covid-19 vaccine over 6 months. *N Engl J Med* 2021;385(24):e84 [FREE Full text] [doi: [10.1056/NEJMoa2114583](https://doi.org/10.1056/NEJMoa2114583)] [Medline: [34614326](https://pubmed.ncbi.nlm.nih.gov/34614326/)]
34. Regev-Yochay G, Gonen T, Gilboa M, Mandelboim M, Indenbaum V, Amit S, et al. Efficacy of a fourth dose of Covid-19 mRNA vaccine against Omicron. *N Engl J Med* 2022;386(14):1377-1380 [FREE Full text] [doi: [10.1056/NEJMc2202542](https://doi.org/10.1056/NEJMc2202542)] [Medline: [35297591](https://pubmed.ncbi.nlm.nih.gov/35297591/)]
35. Atmar RL, Lyke KE, Deming ME, Jackson LA, Branche AR, El Sahly HM, DMID 21-0012 Study Group. Homologous and heterologous Covid-19 booster vaccinations. *N Engl J Med* 2022;386(11):1046-1057 [FREE Full text] [doi: [10.1056/NEJMoa2116414](https://doi.org/10.1056/NEJMoa2116414)] [Medline: [35081293](https://pubmed.ncbi.nlm.nih.gov/35081293/)]
36. Clemens SAC, Weckx L, Clemens R, Mendes AVA, Souza AR, Silveira MBV, RHH-001 study team. Heterologous versus homologous COVID-19 booster vaccination in previous recipients of two doses of CoronaVac COVID-19 vaccine in Brazil (RHH-001): a phase 4, non-inferiority, single blind, randomised study. *Lancet* 2022;399(10324):521-529 [FREE Full text] [doi: [10.1016/S0140-6736\(22\)00094-0](https://doi.org/10.1016/S0140-6736(22)00094-0)] [Medline: [35074136](https://pubmed.ncbi.nlm.nih.gov/35074136/)]
37. Khong KW, Zhang R, Hung IFN. The four Ws of the fourth dose COVID-19 vaccines: why, who, when and what. *Vaccines (Basel)* 2022;10(11):1924 [FREE Full text] [doi: [10.3390/vaccines10111924](https://doi.org/10.3390/vaccines10111924)] [Medline: [36423020](https://pubmed.ncbi.nlm.nih.gov/36423020/)]
38. Arnold DT, Milne A, Samms E, Staddon L, Maskell NA, Hamilton FW. Symptoms after COVID-19 vaccination in patients with persistent symptoms after acute infection: a case series. *Ann Intern Med* 2021;174(9):1334-1336 [FREE Full text] [doi: [10.7326/M21-1976](https://doi.org/10.7326/M21-1976)] [Medline: [34029484](https://pubmed.ncbi.nlm.nih.gov/34029484/)]
39. Wisnivesky JP, Govindarajulu U, Bagiella E, Goswami R, Kale M, Campbell KN, et al. Association of vaccination with the persistence of Post-COVID symptoms. *J Gen Intern Med* 2022;37(7):1748-1753 [FREE Full text] [doi: [10.1007/s11606-022-07465-w](https://doi.org/10.1007/s11606-022-07465-w)] [Medline: [35266128](https://pubmed.ncbi.nlm.nih.gov/35266128/)]
40. Wynberg E, Han AX, Boyd A, van Willigen HDG, Verveen A, Lebbink R, RECOVERED Study Group. The effect of SARS-CoV-2 vaccination on post-acute sequelae of COVID-19 (PASC): a prospective cohort study. *Vaccine* 2022;40(32):4424-4431 [FREE Full text] [doi: [10.1016/j.vaccine.2022.05.090](https://doi.org/10.1016/j.vaccine.2022.05.090)] [Medline: [35725782](https://pubmed.ncbi.nlm.nih.gov/35725782/)]
41. Tang R, Zheng H, Wang BS, Gou JB, Guo XL, Chen XQ, CanSino COVID-19 Study Group. Safety and immunogenicity of aerosolised Ad5-nCoV, intramuscular Ad5-nCoV, or inactivated COVID-19 vaccine CoronaVac given as the second booster following three doses of CoronaVac: a multicentre, open-label, phase 4, randomised trial. *Lancet Respir Med* 2023;11(7):613-623 [FREE Full text] [doi: [10.1016/S2213-2600\(23\)00049-8](https://doi.org/10.1016/S2213-2600(23)00049-8)] [Medline: [36898400](https://pubmed.ncbi.nlm.nih.gov/36898400/)]

Abbreviations

- ELISA:** enzyme-linked immunosorbent assay
- GMI:** geometric mean increase
- GMT:** geometric mean titer
- IgA:** immunoglobulin A
- IgG:** immunoglobulin G
- LOWESS:** locally weighted scatterplot smoothing
- OR:** odds ratio
- PCC:** post-COVID-19 condition
- VOC:** variant of concern

Edited by A Mavragani, T Sanchez; submitted 14.03.23; peer-reviewed by L Rodewald, H Tian; comments to author 08.06.23; revised version received 25.06.23; accepted 08.08.23; published 11.10.23

Please cite as:

Liu M, Zhao T, Mu Q, Zhang R, Liu C, Xu F, Liang L, Zhao L, Zhao S, Cai X, Wang M, Huang N, Feng T, Lei S, Yang G, Cui F
Immune-Boosting Effect of the COVID-19 Vaccine: Real-World Bidirectional Cohort Study
JMIR Public Health Surveill 2023;9:e47272
URL: <https://publichealth.jmir.org/2023/1/e47272>
doi: [10.2196/47272](https://doi.org/10.2196/47272)
PMID:

©Ming Liu, Tianshuo Zhao, Qiuyue Mu, Ruizhi Zhang, Chunting Liu, Fei Xu, Luxiang Liang, Linglu Zhao, Suye Zhao, Xianming Cai, Mingting Wang, Ninghua Huang, Tian Feng, Shiguang Lei, Guanghong Yang, Fuqiang Cui. Originally published in JMIR

Public Health and Surveillance (<https://publichealth.jmir.org>), 11.10.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Public Health and Surveillance, is properly cited. The complete bibliographic information, a link to the original publication on <https://publichealth.jmir.org>, as well as this copyright and license information must be included.

1 Improving Cardiovascular Risk Prediction through Machine Learning 2 Modelling of Irregular Repeated Electronic Health Records

3
4 Chaiquan Li, MS¹, Xiaofei Liu, MS¹, Peng Shen, MD², Yexiang Sun, MPH², Tianjing Zhou, BS¹, Weiye Chen,
5 MS¹, Qi Chen, MS², Hongbo Lin, MD², Xun Tang, PhD, MHS^{1,3*}, Pei Gao, PhD^{1,3,4*}

6
7
8 1. Department of Epidemiology and Biostatistics, School of Public Health, Peking University Health Science Center,
9 2. Yinzhou District Center for Disease Control and Prevention, Ningbo, China,
10 3. Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing, China,
11 4. Center for Real-world Evidence Evaluation, Peking University Clinical Research Institute, Beijing, China,

12
13 *Corresponding Author:

14 Pei Gao: peigao@bjmu.edu.cn; Xun Tang: tangxun@bjmu.edu.cn

17 Lay Summary:

18 The usual cardiovascular risk assessment tools use single measurements of limited traditional risk factors. Existing
19 electronic health records (EHRs) often have abundant longitudinal measurements and a wider range of predictors
20 available. These could not only facilitate the improvement of the prediction accuracy but also allow automatic
21 screening when the tool is embedded within the EHR system. Machine learning approaches are known to
22 accommodate irregular measurement records. This study, therefore, compared the performance of two machine
23 learning models with the guideline-recommended model under real-world scenarios, indicating that:

- 24 • Incorporating irregular multiple predictors with repeated measurements with simple machine learning
25 algorithms was feasible and interpretable.
- 26 • The accuracy of the risk prediction can be significantly improved, especially regarding risk reclassification.
27 According to the risk cut-offs recommended by the current guideline, the machine learning models could
28 allocate the participants into different risk groups more correctly than the guideline-recommended model.

© The Author(s) 2023. Published by Oxford University Press on behalf of the European Society of Cardiology. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27

Abstract

Aims

Existing electronic health records often have abundant but irregular longitudinal measurement risk factors available. We aim to leverage such data to improve the risk prediction of atherosclerotic cardiovascular disease (ASCVD) by applying machine learning algorithms, which can therefore allow the automatic screening of the population.

Methods and results

Totally 215,744 Chinese adults aged 40-79 without a history of CVD from an EHR-based longitudinal cohort study were included (6,081 cases). To allow the model interpretable, predictors of demographic characteristics, medication treatment, and repeatedly measured records of lipids, glycemia, obesity, blood pressure, and renal function were used. The primary outcome was ASCVD, defined as non-fatal acute myocardial infarction, coronary heart disease death, or fatal and non-fatal stroke. The eXtreme Gradient boosting (XGBoost) machine and LASSO regression models were derived to predict the 5-year ASCVD risk. In the validation set, compared with the refitted Chinese guideline-recommended Cox model (i.e., the China-PAR), the XGBoost model had significantly highest C-statistics (0.792, the difference in C-statistics: 0.011, 0.006-0.017, $P<0.001$), with the similar results for LASSO regression (the difference in C-statistics: 0.008, 0.005-0.011, $P<0.001$). The XGBoost model demonstrated the best calibration performance (Men: $D_x=0.598$, $P=0.75$; Women: $D_x=1.867$, $P=0.08$). Moreover, the machine learning algorithms' risk distribution differed from the conventional model. The NRIs of XGBoost and LASSO over the Cox model were 3.9% (1.4%-6.4%) and 2.8% (0.7%-4.9%), respectively.

Conclusions

Machine learning algorithms with irregular, repeated real-world data could improve cardiovascular risk prediction. They demonstrated significantly better performance for reclassification to identify the high-risk population correctly.

Keywords:

Prediction, Preventive Cardiology, Risk

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27

Introduction

Safe and cost-effective treatments can reduce cardiovascular risk significantly. The magnitude of treatment benefit is directly related to the pre-treatment cardiovascular risk of individual patients. To accurately determine this risk, reliable risk prediction equations must be employed. The global cardiovascular guidelines recommend various risk assessment tools to tackle the heavy burden of cardiovascular disease (CVD), e.g., the Pooled Cohort Equation (PCE),¹ the Systematic COronary Risk Evaluation (SCORE) model,² and the Prediction for Atherosclerotic Cardiovascular Disease Risk in China (China-PAR) model.³ Though current risk prediction models using a number of traditional CVD risk factors have played an important role in CVD prevention, the predictive performance has yet to be satisfactory. For example, several external validation studies have demonstrated that the C-statistics of these models were only 0.65 to 0.74,^{4,5} and may incorrectly estimate the absolute cardiovascular risk.^{4,6} It is known that traditional CVD risk prediction can be enhanced through additional information gained from either new predictors or repeat measurements. In addition to traditional predictors such as age, smoking, and systolic blood pressure, new predictors from various etiological pathways (e.g., lipoprotein (a) and apolipoprotein,^{7,8} glucose metabolism,^{9,10} and renal function markers¹¹) could also potentially improve the prediction accuracy. However, implementing traditional prediction models using all these novel predictors for CVD primary prevention in the entire population is not realistic in real-world clinical practice. Secondly, recent evidence suggests that repeated measurements of CVD risk factors in traditional prediction models could improve performance,¹²⁻¹⁴ which may capture the longitudinal information of risk factors and help explain the cardiovascular residual risk.¹⁵ But the current traditional models have limitations in considering a limited number and type of repeated predictors¹⁶, and they may overlook potential interactions among these predictors.^{17,18}

In contrast, electronic health records (EHR) can not only provide a wealth of information with repeat measurements on various predictors¹⁶ but also allow for automated screening if the risk prediction tool is embedded.^{19,20} However, the data structure in real-world EHR systems often differs from that of traditional cohort studies. Although new predictors may exist in subgroups of the population, the pattern of available risk factors is often irregular. For

1 example, patients may have a series of repeated measurements, especially for traditional CVD risk factors, but the
2 number of repeats varies among subjects. Moreover, it is also quite common that different predictors were measured
3 between individuals, even from the same etiological pathway. E.g., someone had information on body mass index,
4 whereas others measured waist-hip ratio. Besides, risk factors were generally measured at different time points.
5 Therefore, this EHR-based information remains challenging to be incorporated using conventional risk prediction
6 models.

7
8 In this case, machine learning (ML) algorithms can be a valuable alternative to handle such complex data. While
9 evidence shows that the benefits of ML algorithms over traditional models using the same predictors were limited, it
10 can excel in accommodating multiple predictors and handling irregular measurements, making them suitable for
11 leveraging the rich information present in EHRs effectively.^{21,22} While ML has been increasingly utilized to leverage
12 information from repeated measurements in certain hospital-based scenarios,^{23,24} its application in primary care for
13 cardiovascular risk assessment remains limited.^{25,26} Existing studies have demonstrated that ML can enhance risk
14 prediction,^{22,27} but they have not fully utilized time-to-event information or comprehensively evaluated predictive
15 performance. Developing fixed-term survival prediction models is crucial for CVD risk assessment, as they align
16 with the recommended risk stratification cut-offs in clinical guidelines.¹⁻³

17
18 Therefore, this study aims to investigate the improvement of CVD risk predictions by incorporating irregular
19 repeated real-world measurements of multiple predictors using ML models. The predictive performance was then
20 compared against the guideline-recommended traditional Cox regression model.²⁸

22 **Methods**

23 **Study design**

24 The concept of the study design is shown in **Figure 1-(a)**. The population included in this study was from the
25 CHinese Electronic health Records Research in Yinzhou (CHERRY) study, which was an EHR-based cohort study
26 in Yinzhou, Ningbo (a developed area in Eastern China). A detailed description of the CHERRY study has been

1 published elsewhere.²⁹ The inclusion criteria of this study population consisted of 1) aged between 40 to 79 years
2 old at the entry time; 2) registered in the health information system from Jan 1st, 2010 to Dec 31st, 2016; and 3)
3 Chinese residents who had been living in Yinzhou for at least six months. The exclusion criteria of this study are as
4 follows: 1) had no records of serum lipids measurements since lipid-related predictors were causally related to
5 atherosclerotic cardiovascular disease (ASCVD); This may cause the model to be not applicable. 2) had
6 cardiovascular disease history before entering the study. The flowchart of the inclusion and exclusion process is
7 shown in **Supplementary Figure 1**. Finally, 215,744 participants were included in the analysis set, among which a
8 random sample of 80% (about 180,000) was separated as the training set to derive the models, and the rest
9 participants were left only for the final internal validation (Shown in **Supplementary Figure 2**). This study was
10 approved by the Peking University Institutional Review Board (IRB00001052-16011).

11
12 To maximize the number of repeated measurements collected, the baseline in this study was set as 1) the time when
13 the participants registered in the system, 2) the time when participants reached 40 years old, 3) the time when the
14 first serum lipids measurement was recorded, or 4) Jan 1st, 2015, whichever the latest. The repeated measurements
15 were collected from the past five years before the baseline. Participants would be followed up to the time 1) when
16 they had their first ASCVD event (further defined in the *outcomes* section), 2) they were censored from following
17 up, or 3) May 31st, 2020, whichever is the earliest.

19 **Predictors**

20 Seven common categories of cardiovascular risk factors (Shown in **Figure 1-(b)**) with 25 markers in total were pre-
21 identified as the pool of predictors, including demography (age, sex, education levels, settings, smoke status, and
22 family history), lipid metabolism (total cholesterol [TC], high-density lipoprotein cholesterol [HDL-C], low-density
23 lipoprotein cholesterol [LDL-C], triglycerides [TG], apolipoprotein A [apo A], apolipoprotein B [apo B], and
24 lipoprotein (a) [Lp-(a)]), obesity (body mass index [BMI] and waist circumference), glucose metabolism (fasting
25 blood glucose [FBG], diabetes at baseline, and hemoglobin A1c [HbA1c]), blood pressure (systolic blood pressure
26 [SBP] and diastolic blood pressure [DBP]), renal function (estimated glomerular filtration rate [eGFR] and albumin
27 creatinine ratio [ACR]), and medical treatments (antihypertension, antihyperglycemic, antihyperlipidemic treatment,
28 and aspirin). We selected these risk factors because they were universally incorporated into cardiovascular risk

1 prediction,^{1,12,28,30-32} had likely causal relationships with ASCVD outcomes,^{8,9,33,34} or were closely associated with
2 ASCVD from etiological perspectives.^{7,10,11,35,36} Measurements of these predictors were collected from multiple
3 sources in the regional health system, including census data, electronic medical records (EMR), disease surveillance,
4 chronic disease management system, and health check, etc., which were summarized in **Supplementary Table 1**.
5 These records will be inherently linked to each other according to a unique and encoded identifier. Detailed data
6 collection procedures of various data sources were described in **Supplementary Method 1**. The exact definitions of
7 each medical treatment are given in **Supplementary Table 2**. Extreme outliers were removed according to pre-
8 specified normal ranges of key predictors (Shown in **Supplementary Table 3**).

9
10 Considering the irregular nature of the predictors' information available, we used a simple but effective approach to
11 leverage these repeated measurements by summarized statistics.^{24,37} Standard deviation, range, and the difference
12 between the last and first measurements were calculated as derived predictors since many studies proposed that the
13 variability of predictors was associated with CVD.^{17,18,38} The numbers of measurements were also counted and
14 included in the pool of predictors.³⁹ Mean values of predictors were also summarized to represent the long-term
15 average of those predictors. All the baseline and derived predictors included in this research were listed in
16 **Supplementary Table 4**.

18 **Outcomes**

19 The definition of the ASCVD was consistent with the one used in the China-PAR or PCE model, which was defined
20 as the composite outcome of non-fatal or fatal stroke (ICD-10 code: I60, I61, I63, I64), non-fatal myocardial
21 infarction (I21, I22), and coronary heart disease death (I20-25).²⁸ Outcomes in this study were collected from the
22 following sources: disease surveillance, chronic disease management system, death registry, and EMR. Among those
23 sources, the disease surveillance and death registry were recognized as the gold standard. The outcome used the first
24 ASCVD events that occurred after the baseline and before May 31st, 2020.

26 **Risk prediction models**

27 Since the China-PAR model was the Chinese guideline-recommended risk assessment tool in primary care, our
28 study selected this model as the reference to be compared. The model was modified by two different approaches in

1 this study to make the comparison fair: (1) the refitted China-PAR model was developed by directly replacing all the
2 coefficients in the original model but preserving all the pre-defined terms (including all the interaction terms); (2)
3 the recalibrated China-PAR model was developed by replacing the baseline survivals and means of linear predictors
4 in the original model without altering any pre-defined terms and their corresponding coefficients.

5
6 Two ML approaches were finally adopted in this study, which were eXtreme Gradient Boosting (XGBoost) and
7 Least Absolute Shrinkage and Selection Operator (LASSO) regression. The choice of algorithms depends on various
8 factors, including the nature of the data, the size of the dataset, the complexity of the problem, and the desired
9 interpretability of the model.¹⁹ For large datasets with high dimensionality (many predictors), algorithms that can
10 efficiently handle such data, like Random Forest, Gradient Boosting, or Deep Learning models, may be suitable.^{40,41}
11 Meanwhile, for CVD risk prediction, model interpretability is crucial. Simpler models such as regression-based
12 models or decision trees are still preferred, as they can provide more transparent and easily interpretable results.¹⁹
13 Random Forest or Gradient Boosting can also offer feature importance rankings and handle missing values without
14 imputation. Finally, the computational cost of training the model is a consideration, especially for large datasets.
15 Linear models and tree-based models tend to be faster to train compared to deep learning models. To control the
16 potential overfitting, algorithms with built-in regularization, such as Lasso Regression was considered. After
17 exploring the performance and feasibility of the four aforementioned methods, XGBoost and LASSO regression
18 were chosen because they have better performance and are relatively readily to be interpreted and implemented in
19 the EHR system. The difference between the perspectives these two algorithms utilize the information of predictors
20 will also provide a comprehensive exploration of the suitable approach to leverage the repeated measurements.^{42,43}
21 The importance of predictors was assessed according to the average reduction of information entropy in the
22 XGBoost model and the absolute value of the β coefficients in LASSO regression, which reflected the information
23 gains or the marginal effects of predictors. Two ML classifiers were first trained in the 126,893 subjects with known
24 outcome information at the end of the fifth year. Then, the two ML models were embedded into a Cox regression
25 model to predict absolute 5-year risk. Hyperparameter tuning was conducted by maximizing the area under curve in
26 the five-fold cross-validation. Grid search were iterated 100 times to acquire the optimized hyperparameter
27 sequence.^{44,45} The ranges of hyperparameters were given in **Supplementary Table 5**.

28

1 **Statistical analysis**

2 Continuous predictors were described using means and standard deviations, while categorical predictors were
3 described using counts and percentages. The associations between predictors and ASCVD were given according to
4 the hazard ratios of Cox proportional hazard regression adjusted for the variables from the China-PAR models.
5 Proportions of missingness were described for each predictor. The predictors in the China-PAR models were
6 multiple-imputed by chain equations (MICE, five imputation sets were created) to compare with the two ML
7 models.^{12,46} The performance metrics were measured in each imputation set and then pooled according to Rubin's
8 rules.⁴⁷ ML models can handle the missingness directly to preserve initial information. Details on data imputation
9 were described in **Supplementary Method 2**. The performances of the models were evaluated from the following
10 perspectives, discrimination, calibration, and reclassification. The discrimination was assessed by Harrell's C-
11 statistics. Calibration was used to measure the coordination between predicted risk and observed risk, which was
12 evaluated using the Hosmer-Lemeshow χ^2 and calibration plots.⁴⁸ C-statistics of different models were compared
13 using the approach proposed by Kang et al.⁴⁹ The risk distribution by different models was also illustrated. To pool
14 the Hosmer-Lemeshow χ^2 given by different imputation sets, a D_x statistic following the F distribution was
15 generated based on the approach proposed by Rubin et al.⁵⁰ and Li et al.⁵¹ We provided the standard reclassification
16 table. Net reclassification improvement (NRI) and integrated discrimination improvement (IDI) were calculated to
17 quantify the reclassification benefits of the ML models over the refitted China-PAR model. The cut-offs of risk
18 groups were selected according to the 2019 Guideline on the assessment and management of cardiovascular risk in
19 China.³ A decision curve analysis was also conducted to illustrate the clinical implications of the ML models. The
20 sensitivity analyses were conducted as follows: (1) to further ascertain whether the possible improvement of risk
21 prediction was driven by leveraging the information from the repeated measurements or by simply including more
22 baseline predictors, a Cox regression model including all the baseline measurements of each predictor was
23 constructed, and its performance was compared against the two ML models and the refitted China-PAR model; (2)
24 the performance of recalibrated China-PAR was assessed and compared to evaluate how much improvement ML
25 models can achieve compared with the per-guideline approach. All analyses were conducted using R version 4.0.4
26 with a statistical significance level of $P < 0.05$. The XGBoost model was constructed with the *xgboost* package
27 version 1.4.1.1, and the LASSO regression was conducted using *glmnet* package version 4.1-1. The *mice* package
28 version 3.13.0 was adopted for the multiple-imputation by chain equations.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27

Results

Basic descriptions for participants and predictors

The characteristics of the 215,744 included participants were described in **Table 1**. Fifty-four percent of the participants were women, and the mean age was about 56.7 (SD = 9.6). The means of major risk factors for ASCVD: systolic blood pressure, total cholesterol, and high-density lipoprotein cholesterol were 134.5 mmHg, 4.9 mg/dL, and 1.3 mg/dL, respectively. The average BMI was 23.3 kg/m². Overall, 12.1% of them had diabetes at baseline. During a median of 5.4-year follow-up, 6081 individuals (2.82%) had ASCVD outcomes. The incidence rate of ASCVD was 6,178 per million person-years. Only total cholesterol and anti-hyperglycemia treatment significantly differed between derivation and internal validation datasets (Shown in **Supplementary Table 6**). The missing proportions of each predictor were shown in **Supplementary Table 7**. The number of measurements and time intervals between each measurement of key predictors for each individual was given in **Supplementary Table 8**. The mean number of measurements of total cholesterol, SBP, BMI, and fasting glucose were 3, 2, 1, and 3, respectively. The corresponding median time intervals between those measurements were 269, 136, 267, and 251 days.

The discrimination of the models

In the validation set, the C statistics with the absolute differences compared with refitted China-PAR model were shown in **Figure 2**. The C statistics of the XGBoost model were 0.7918 (95% CI: 0.7776-0.8060) and that of LASSO regression was 0.7883 (0.7737-0.8029). The two ML models performed better than the refitted China-PAR model in discrimination (Difference in C statistics for XGBoost: 0.01134, 0.00567-0.01700, $P < 0.001$; for LASSO: 0.00784, 0.00453-0.01115, $P < 0.001$). The discrimination of the two ML models was better than the refitted China-PAR model in both men and women, where the XGBoost model performed the best among men while the LASSO regression performed the best among women. The final hyperparameters in the final ML models were in **Supplementary Table 9**. The major structures of the final ML models were given in **Supplementary Figure 3** and **Supplementary Table 10**.

1 **The calibration of the models**

2 XGBoost model showed better calibration than refitted China-PAR model in both men and women (XGBoost: $D_x =$
3 $0.598, P = 0.75$ in men and $D_x = 1.867, P = 0.08$ in women; Refitted China-PAR: $D_x = 2.832$ in men, $P = 0.004$ and
4 $D_x = 3.352$ in women, $P = 0.001$) while LASSO regression was recalibrated well in men ($D_x = 1.639, P = 0.11$) but
5 not in women ($D_x = 1.950, P = 0.048$). The calibration plots were shown in **Figure 3**. Although the XGBoost model
6 slightly overestimated the risk in the highest risk group, the coordination of the predicted risks and Kaplan-Meier
7 observed risks was much better than the LASSO model and refitted the China-PAR model, especially among low-
8 risk deciles.

10 **The clinical implications on the outcomes**

11 In the validation set, the reclassification table was shown in **Table 2**. By using the XGBoost model compared with
12 the refitted China-PAR model, among 24,247 non-case individuals, there were 3,355 and 667 subjects classified into
13 low or medium-risk by the XGBoost and the refitted China-PAR model respectively. A net quantity of 2,688 people
14 (11.09%) was reclassified into the correct groups. Among 969 individuals who developed CVD during follow-up,
15 XGBoost, and the refitted China-PAR model selected a similar number of high-risk subjects, i.e., 550 and 585. After
16 taking the medium-risk group into account, China-PAR correctly selected 70 (7.22%) more case individuals. The
17 overall net reclassification improvement (NRI) is 3.87% (1.35-6.38%). Similarly, the NRI for LASSO regression is
18 2.78% (0.66-4.89%). By directly comparing the predicted risk against the refitted China-PAR model, the integrated
19 discrimination improvements (IDI) of the XGBoost model and LASSO regression were 0.0174 (0.0135-0.0212) and
20 0.0106 (0.0081-0.0131). The risk distributions predicted by the XGBoost model and the refitted China-PAR model
21 were illustrated in **Figure 4**. The risk predicted by XGBoost tended to centralize in the lower range in non-cases in
22 both men and women, with a larger difference between the risks of cases and non-cases. The decision curve analysis
23 (DCA) demonstrated that all three models, namely XGBoost, LASSO, and the refitted China-PAR model, exhibited
24 favorable performance by deviating from the curves of treating all or treating none within the common
25 cardiovascular risk range of 0%-20%. Moreover, the net benefit of the XGBoost model surpassed that of the refitted
26 China-PAR model between the threshold range of 7.5% and 12.5%, while the net benefit of the LASSO regression
27 model was superior within the range of 12.5% to 17.5% (**Supplementary Figure 4**).

28

1 **The importance of predictors**

2 The associations between the predictors and ASCVD were presented in **Supplementary Table 11** adjusted by
3 predictors in the China-PAR model. All the predictors were included in the XGBoost model because of its random
4 subspace sampling, while the LASSO regression selected only 78 of the total 101 predictors (i.e., baseline and
5 summarized statistics of repeat information of 25 markers). The rank of importance was given in **Supplementary**
6 **Figure 5**. In general, age, anti-hypertension treatment history, glucose metabolism-related predictors, lipid
7 metabolism-related predictors, blood pressure, eGFR, and family history of ASCVD were most valued by both ML
8 models. The importance of fasting blood glucose ranked third and fifth in the XGBoost model and LASSO
9 regression, respectively. The novel lipid predictor, such as Apo B, ranked eighth and tenth in the two ML models,
10 while the classic predictor, like total cholesterol, ranked only tenth and seventeenth. The importance of smoking and
11 predictors indicating obesity were relatively lower (BMI: 16 in XGBoost and 19 in LASSO; Waist circumference:
12 20 in XGBoost and 16 in LASSO; smoking: 18 in XGBoost and 14 in LASSO).

14 **Sensitivity analysis**

15 The Cox regression model with baseline measurements of all the predictors performed better than the refitted China -
16 PAR model, while it was still worse than the XGBoost model in the whole validation set from the perspective of
17 discrimination (The differences of C statistics: 0.00563, 0.00118-0.01009, $P = 0.01$, **Supplementary Table 12**). Its
18 discriminative performance was not significantly different from the LASSO regression (0.00214, -0.00088-0.00515,
19 $P = 0.17$, **Supplementary Table 12**). The calibration plot of the Cox model with all the baseline measurements was
20 not coordinated enough compared with the two ML models, which was not even better than the refitted China -PAR
21 model ($D_x = 2.421$, $P = 0.01$ in men and $D_x = 2.216$, $P = 0.02$ in women, **Supplementary Figure 6**). Both of the ML
22 models performed significantly better than the recalibrated China -PAR model no matter in discrimination (all $P <$
23 0.001, **Supplementary Table 13**) or in calibration (Recalibrated China-PAR: $D_x = 2.421$ in men and $D_x = 2.216$ in
24 women, both $P < 0.001$, **Supplementary Figure 6**).

25
26

1 Discussion

2 This study used two ML approaches (XGBoost and LASSO) to leverage the existing repeated measurements in EHR
3 data to predict 5-year atherosclerotic cardiovascular risk. Both ML models outperformed the recalibrated and even
4 the refitted China-PAR model from the perspectives of discrimination, calibration, and reclassification, which is the
5 model recommended by the 2019 Guideline on the assessment and management of cardiovascular risk in China.³
6
7 Repeated measurements from electronic health records (EHR) offer valuable contributions to cardiovascular risk
8 prediction. Notably, the QRISK3 model in the UK was derived from EHR data obtained from general practices'
9 computer systems, where the standard deviation of SBP was included as a predictor¹². This model stands as the first
10 nationwide-used risk prediction model to incorporate predictors derived from repeated blood pressure
11 measurements. Similarly, Paige et al. leveraged EHR data from the Health Improvement Network, a United
12 Kingdom general practice electronic database, and applied a landmark model to utilize information from repeated
13 measurements of smoking status, SBP, TC, and HDL-C, resulting in a significant improvement in C-statistic¹⁴. Our
14 study aligns with these findings, demonstrating that incorporating repeated measurements of multiple predictors
15 from EHRs enhances predictive performance when compared to the Cox model that only uses baseline
16 measurements. The temporal information present in repeated measurements is of great importance. It was reflected
17 by the time intervals between measurements and the trends or patterns observed over time. While the QRResearch
18 study and Paige et al. did not explicitly report the time intervals (or density) of measurements, Paige et al. did fit the
19 temporal trend and dependency of repeated measurements using a multivariate linear mixed-effects model.^{12,14} In
20 our study, we observed that the average time intervals between measurements of key predictors were generally less
21 than one year (**Supplementary Table 8**), signifying the richness of information that can be harnessed from EHRs.
22 The correlated predictors,²² irregularly-missing records,⁵² and data with strong interaction in the EHR necessitate
23 applying a novel modeling approach such as ML which usually utilizes high-dimensional unstructured data to
24 enhance the predictive performance.^{53,54} However, it is worth noting that existing ML algorithms lack a
25 comprehensive approach to model secular trends and dependencies in irregularly structured data. This presents an
26 area for further methodological investigation to effectively harness the temporal nature of the data for CVD risk
27 prediction.

1
2 Although it is controversial whether ML can improve cardiovascular risk prediction using only baseline
3 measurements of limited predictors,^{27,55} several pieces of evidence demonstrated that predictive performance could
4 be largely improved when predictors derived from repeated measurements were fed into ML models.^{25,26} For
5 instance, Li et al. summarized the repeated measurements of blood lipid, blood pressure, and HbA1c from the EMRs
6 of 101,110 people in a US regional healthcare system, into extremum, number of measurements, and means, etc.
7 Then, these longitudinal-derived predictors were used in the random forest ML model, causing large increments of
8 AUC (e.g., 0.823 to 0.902).²⁵ Compared to those studies, our study demonstrated that: (1) By embedding XGBoost
9 and LASSO regression algorithms into the Cox regression to leverage the time-to-event information, we found
10 similar improvement in discrimination when evaluated by C statistics; (2) Besides the discrimination capability, our
11 study comprehensively assessed the performance of the model from the perspectives of calibration and
12 reclassification based on survival probabilities; (3) It is feasible to conduct CVD risk prediction using rich but
13 irregular existing EHR data for risk stratification without extra cost for screening new markers.

14
15 Under real-world scenarios, many predictors are not universally screened in the population. However, it was shown
16 that these markers can predict cardiovascular risk. For example, the mean of fasting blood glucose presents the long-
17 term control of glucose metabolism, which was predictive for cardiovascular disease independently.¹⁰
18 Apolipoprotein B and Lp (a) are also useful biomarkers for ASCVD.⁷ Poor renal function (e.g., impaired eGFR)
19 could result in hypertension, left ventricular hypertrophy, endothelial dysfunction, dyslipidemia, and low-grade
20 inflammation.⁵⁶ In our study, these predictors were informative in predicting cardiovascular events as reflected by
21 the importance of predictors (**Supplementary Figure 5**) and the structures of the models (**Supplementary Figure 3**
22 and **Supplementary Table 10**). Making the best use of these existing biomarkers in EHR data to enhance CVD risk
23 prediction may change the current way of screening high-risk populations in clinical practice. Considering the
24 irregular nature of the data, ML algorithms can be good alternatives. The ML models could accommodate residents
25 with some unmeasured predictors flexibly. Including a predictor or its repeated measurement in the model does not
26 necessitate requiring complete information on the whole population.

27
28 The absolute increment of C-statistics in our study was 0.0113. This gain in discrimination was meaningful

1 compared to the gains generated by established risk factors. For illustration, in the Emerging Risk Factors
2 Collaboration study, adding C reaction protein or HLD-C into the traditional Cox model to predict ASCVD incidents
3 will increase the C-statistics by 0.0039 or 0.0050, respectively.⁵⁷ When SBP was removed from the Reynolds score
4 in the Women's Health Study, the change of C-statistic was 0.01.⁵⁸ C-statistic was an insensitive indicator that
5 ranges from 0.5 to 1.0. The larger the C-statistic is, the more challenging for it to be improved.⁵⁹ HDL-C could only
6 increase the C-statistic by 0.0013 in our cohort. Advised by Cook,⁵⁸ the improvement of risk prediction given by the
7 two ML models was also evaluated using NRI and IDI in this study. The reclassification table of the XGBoost model
8 indeed indicated significant net benefit. In the validation set with 25,216 subjects, according to the cut-offs defined
9 by the current Chinese guideline, about 4% more subjects will be allocated to proper risk groups and
10 correspondingly receive more suitable recommendations on intervention. Assuming that the statin therapy was
11 recommended to the high-risk population and reduced the CVD risk by 20%,⁶⁰ such assessments of individuals by
12 the XGBoost and the refitted China-PAR model could assign 4529 (17.9%) and 5398 (21.4%) patients to initiate the
13 statin treatment and help prevent 110 and 117 CVD outcome over 5 years respectively. Correspondingly for every 41
14 and 46 patients treated, there was 1 CVD outcome prevented by using the XGBoost and the refitted China-PAR
15 model. This is consistent with the calibration plot where the risk predicted by the XGBoost model was more
16 coordinated to the observed risk than refitted China-PAR model, especially among the low or intermediate-risk
17 groups. Such consistency indicates the XGBoost model may gain the benefit under the existing risk cut-off values.
18 Considering the large number of the low-risk population, great benefits are likely to be achieved when this model is
19 implemented for risk screening. In the decision curve analysis, the threshold probability defined the criteria for
20 intervention in individuals. If the estimated risk exceeded the threshold probability, intervention would be
21 recommended. The net benefit of the XGBoost model outperformed the refitted China-PAR model within the
22 threshold probability range of 7.5% to 12.5%. This range aligns with the typical cut-off risk values recommended by
23 guidelines for initiating critical important interventions, such as statin therapy.^{1,2} These results suggest the potential
24 net benefit of implementing the XGBoost model based on the existing risk cut-offs. On the other hand, the risk
25 predicted by the LASSO regression may be more suitable for use in high-risk individuals, given its larger net benefit
26 across the range of 12.5% to 17.5%. Finally, we note that this administrative-data-based approach can enhance CVD
27 primary prevention by offering a more accurate prediction without any extra cost for screening new markers.

28

1 In the present landscape, most risk prediction models have developed their own implementation tools, some of
2 which are integrated into the health information system (e.g., QRISK in the UK and PREDICT in New Zealand),^{6,12}
3 while others are offered independently through websites or applications (e.g., PCE, SCORE2, and the China-PAR
4 model).^{1,28,30} Given the nature of utilizing comprehensive information from electronic health records (EHRs), we
5 recommend implementing the machine learning model by embedding it within the healthcare information system.
6 This approach can also facilitate automatic population screening, enhancing the sustainability of cardiovascular risk
7 prediction. However, unlike the traditional Cox model, the implementation of an already derived ML model is not
8 always straightforward. Our algorithm for 5-year prediction of CVD risk involved a two-stage process. The first
9 stage utilized ML classification algorithms, while the second stage incorporated the ML classifier into a Cox
10 regression model to predict absolute 5-year risk. Therefore, we firmly believe that the baseline survival
11 characteristics of local populations remain crucial for accurate absolute risk prediction. As a result, recalibration of
12 the model may still be necessary when applying it to different populations, along with external validation to assess
13 its performance in diverse settings.

14
15 Our study also has several limitations. First, though internally validated, the ML risk prediction models derived in
16 our study were not externally and independently validated. Our study aims not to propose and generalize the ML
17 models to other populations but to answer a methodological question by comparing the performance of two ML
18 approaches to the locally refitted China-PAR models. The models' relative performance was still valid since the
19 performance was all measured by the same scale from the same dataset. Secondly, only two ML methods were
20 present in this study, considering the nature and sample size of the data, the complexity of the algorithm, and the
21 desired model interpretability. Advanced ML methods such as neural networks, can be adapted to use the data in the
22 future.⁶¹ Thirdly, our study is based on regional data, which may not fully represent the diversity of the Chinese
23 population nationwide. Variations in genetic background, culture, socioeconomic levels, climate, geographic
24 features, lifestyle, and dietary patterns among different ethnic groups within the Chinese population could influence
25 the generalizability of our findings. Nevertheless, the primary objective of our study was to demonstrate the
26 cardiovascular predictive value of repeated measurements using machine learning models. As such, the potential
27 limitations arising from regional data may have a limited impact on the overall conclusions of this research.

28 Additionally, we acknowledge that the analysis set, consisting of 215,744 Chinese participants, is a subset of the

1 original CHERRY study, which included 1.05 million adults. Consequently, while our findings are informative, they
 2 may not fully represent the entire population. Nonetheless, this subset reflects the current clinical practice where
 3 lipid measurements are commonly requested, even when using traditional guideline-recommended models.
 4 Furthermore, it is important to note that the data source for our study primarily relied on EHRs, which are generally
 5 collected from individuals seeking medical care. This approach may lead to biased representations of certain health
 6 conditions or risk factors that are more likely to be captured in clinical settings. Novel risk factors, such as
 7 apolipoproteins or eGFR, may be particularly affected by this bias, as their availability could be associated with
 8 specific patient health conditions and outcomes. However, we mitigated this concern by leveraging machine learning
 9 algorithms, which effectively handle missing data and enable us to capture valuable information for CVD risk
 10 prediction, including the association between the availability of specific markers and disease outcomes. Finally,
 11 although using summarized statistics to utilize repeated measurements is common, it is also important to model the
 12 time trend and consider the temporal dependence of the measurements from a single individual.^{16,24} Our study
 13 reinforces the importance of incorporating repeated measurements from EHRs in CVD risk prediction. The temporal
 14 aspect of repeated measurements adds valuable insights, but challenges remain in fully capturing this information
 15 using current ML algorithms. Future research efforts should focus on addressing these methodological limitations to
 16 unlock the full potential of EHR data for improved CVD risk assessment. While our study has several limitations
 17 listed above, we believe that our focus on assessing the cardiovascular predictive value of repeated measurements
 18 with machine learning models remains valuable and contributes to the current understanding of CVD risk
 19 assessment.

20
 21 In conclusion, the irregular repeated measurements in the EHR could be leveraged to improve the current 5-year
 22 ASCVD incident risk prediction by adopting the XGBoost or LASSO regression algorithms. XGBoost model had
 23 the best overall performance from the perspectives of discrimination, calibration, and reclassification.
 24 Comprehensively considering the importance of the predictors in both ML models, the average level of blood
 25 glucose, renal function, and Apo B had relatively higher predictive values. Real-world repeated measurements of
 26 risk factors have the potential to provide additive value for current ASCVD risk assessment.

27
 28

1 **Funding**

2 This work was supported by the National Key Research and Development Program of China (grant number
3 2020YFC2003503) and the National Natural Science Foundation of China (NSFC) (grant number: 81973132).

6 **Acknowledgment**

7 The authors thank Yinzhou District Health Bureau for providing access to the administrative databases used in the
8 study and all the CHERRY study investigators for their contributions.

11 **Conflict of interest**

12 Dr Gao reported receiving research funds from Bayer and Merck. These funding sources had no relation to this
13 study. All other authors have reported that they have no relationships relevant to the contents of this paper to
14 disclose.

17 **Data availability**

18 Data in this study were not publicly available due to administrative management. Any request please contact the
19 author directly.

21

1 Reference

- 2
3 1. Arnett DK, Blumenthal RS, Albert MA, *et al.* 2019 ACC/AHA guideline on the primary prevention of
4 cardiovascular disease: a report of the American College of Cardiology/American Heart Association Task Force on
5 Clinical Practice Guidelines. *Journal of the American College of Cardiology* 2019;**74**:e177-e232.
- 6 2. Visseren FL, Mach F, Smulders YM, *et al.* 2021 ESC Guidelines on cardiovascular disease prevention in
7 clinical practice: Developed by the Task Force for cardiovascular disease prevention in clinical practice with
8 representatives of the European Society of Cardiology and 12 medical societies With the special contribution of the
9 European Association of Preventive Cardiology (EAPC). *European Heart Journal* 2021;**42**:3227-3337.
- 10 3. Gu D. Guideline on the assessment and management of cardiovascular risk in China. *Chin J Prev Med*
11 2019;**53**:13-34.
- 12 4. Kist JM, Vos RC, Mairuhu AT, *et al.* SCORE2 cardiovascular risk prediction models in an ethnic and
13 socioeconomic diverse population in the Netherlands: an external validation study. *Eclinicalmedicine* 2023;**57**.
- 14 5. Muntner P, Colantonio LD, Cushman M, *et al.* Validation of the atherosclerotic cardiovascular disease
15 Pooled Cohort risk equations. *Jama* 2014;**311**:1406-1415.
- 16 6. Pylypchuk R, Wells S, Kerr A, *et al.* Cardiovascular disease risk prediction equations in 400 000 primary
17 care patients in New Zealand: a derivation and validation study. *The Lancet* 2018;**391**:1897-1907.
- 18 7. Mehta A, Shapiro MD. Apolipoproteins in vascular biology and atherosclerotic disease. *Nat Rev Cardiol*
19 2022;**19**:168-179. doi: 10.1038/s41569-021-00613-5
- 20 8. Nordestgaard BG, Chapman MJ, Ray K, *et al.* Lipoprotein (a) as a cardiovascular risk factor: current status.
21 *European heart journal* 2010;**31**:2844-2853.
- 22 9. Au Yeung SL, Luo S, Schooling CM. The impact of glycated hemoglobin (HbA1c) on cardiovascular
23 disease risk: a Mendelian randomization study using UK Biobank. *Diabetes Care* 2018;**41**:1991-1997.
- 24 10. Emergency Risk Factor Collaboration. Diabetes mellitus, fasting blood glucose concentration, and risk of
25 vascular disease: a collaborative meta-analysis of 102 prospective studies. *The Lancet* 2010;**375**:2215-2222.
- 26 11. Lim CC, Teo BW, Ong PG, *et al.* Chronic kidney disease, cardiovascular disease and mortality: a
27 prospective cohort study in a multi-ethnic Asian population. *European journal of preventive cardiology*
28 2015;**22**:1018-1026.

- 1 24. Goldstein BA, Pomann GM, Winkelmayr WC, Pencina MJ. A comparison of risk prediction methods
2 using repeated observations: an application to electronic health records for hemodialysis. *Statistics in medicine*
3 2017;**36**:2750-2763.
- 4 25. Li Q, Campan A, Ren A, Eid WE. Automating and improving cardiovascular disease prediction using
5 Machine learning and EMR data features from a regional healthcare system. *Int J Med Inform* 2022;**163**:104786.
- 6 26. Zhao J, Feng Q, Wu P, *et al.* Learning from Longitudinal Data in Electronic Health Record and Genetic
7 Data to Improve Cardiovascular Event Prediction. *Sci Rep* 2019;**9**:717.
- 8 27. Kakadiaris IA, Vrigkas M, Yen AA, *et al.* Machine learning outperforms ACC/AHA CVD risk calculator
9 in MESA. *Journal of the American Heart Association* 2018;**7**:e009476.
- 10 28. Yang X, Li J, Hu D, *et al.* Predicting the 10-year risks of atherosclerotic cardiovascular disease in Chinese
11 population: the China-PAR Project (Prediction for ASCVD Risk in China). *Circulation* 2016;**134**:1430-1440.
- 12 29. Lin H, Tang X, Shen P, *et al.* Using big data to improve cardiovascular care and outcomes in China: a
13 protocol for the CHinese Electronic health Records Research in Yinzhou (CHERRY) Study. *BMJ open*
14 2018;**8**:e019698.
- 15 30. SCORE2 working group and ESC Cardiovascular risk collaboration. SCORE2 risk prediction algorithms:
16 new models to estimate 10-year risk of cardiovascular disease in Europe. *European heart journal* 2021;**42**:2439-
17 2454.
- 18 31. D'Agostino Sr RB, Vasan RS, Pencina MJ, *et al.* General cardiovascular risk profile for use in primary
19 care: the Framingham Heart Study. *Circulation* 2008;**117**:743-753.
- 20 32. Kaptoge S, Pennells L, De Bacquer D, *et al.* World Health Organization cardiovascular disease risk charts:
21 revised models to estimate risk in 21 global regions. *The Lancet Global Health* 2019;**7**:e1332-e1345.
- 22 33. Grundy SM, Stone NJ, Bailey AL, *et al.* 2018
23 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of
24 Blood Cholesterol: A Report of the American College of Cardiology/American Heart Association Task Force on
25 Clinical Practice Guidelines. *Journal of the American College of Cardiology* 2019;**73**:e285-e350.
- 26 34. Chan II, Kwok MK, Schooling CM. The total and direct effects of systolic and diastolic blood pressure on
27 cardiovascular disease and longevity using Mendelian randomisation. *Scientific Reports* 2021;**11**:1-9.

- 1 35. Liu K, Cedres LB, Stamler J, *et al.* Relationship of education to major risk factors and death from coronary
2 heart disease, cardiovascular diseases and all causes, Findings of three Chicago epidemiologic studies. *Circulation*
3 1982;**66**:1308-1314.
- 4 36. Duran EK, Aday AW, Cook NR, *et al.* Triglyceride-rich lipoprotein cholesterol, small dense LDL
5 cholesterol, and incident cardiovascular disease. *Journal of the American College of Cardiology* 2020;**75**:2122-
6 2135.
- 7 37. Plate JD, van de Leur RR, Leenen LP, *et al.* Incorporating repeated measurements into prediction models in
8 the critical care setting: a framework, systematic review and meta-analysis. *BMC medical research methodology*
9 2019;**19**:1-11.
- 10 38. Stevens SL, Wood S, Koshiaris C, *et al.* Blood pressure variability and cardiovascular disease: systematic
11 review and meta-analysis. *bmj* 2016;**354**.
- 12 39. Goldstein BA, Bhavsar NA, Phelan M, Pencina MJ. Controlling for informed presence bias due to the
13 number of health encounters in an electronic health record. *American journal of epidemiology* 2016;**184**:847-855.
- 14 40. Ambale-Venkatesh B, Yang X, Wu CO, *et al.* Cardiovascular Event Prediction by Machine Learning.
15 *Circulation Research* 2017;**121**:1092-1101.
- 16 41. Hoogeveen RM, Pereira JPB, Nurmohamed NS, *et al.* Improved cardiovascular risk prediction using
17 targeted plasma proteomics in primary prevention. *Eur Heart J* 2020;**41**:3998-4007.
- 18 42. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd*
19 *international conference on knowledge discovery and data mining*. 2016, p.785-794.
- 20 43. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society:*
21 *Series B (Methodological)* 1996;**58**:267-288.
- 22 44. Al-Zaiti SS, Alghwiri AA, Hu X, *et al.* A clinician's guide to understanding and critically appraising
23 machine learning studies: a checklist for Ruling Out Bias Using Standard Tools in Machine Learning (ROBUST-
24 ML). *European Heart Journal - Digital Health* 2022;**3**:125-140.
- 25 45. Mathioudakis NN, Abusamaan MS, Shakarchi AF, *et al.* Development and validation of a machine learning
26 model to predict near-term risk of iatrogenic hypoglycemia in hospitalized patients. *JAMA Network Open*
27 2021;**4**:e2030913-e2030913.

- 1 46. Harel O, Mitchell EM, Perkins NJ, *et al.* Multiple imputation for incomplete data in epidemiologic studies.
2 *American journal of epidemiology* 2018;**187**:576-584.
- 3 47. Rubin DB. Multiple imputation for nonresponse in surveys. Hoboken, N.J: Wiley-Interscience; 2004.
- 4 48. Alba AC, Agoritsas T, Walsh M, *et al.* Discrimination and calibration of clinical prediction models: users'
5 guides to the medical literature. *Jama* 2017;**318**:1377-1384.
- 6 49. Kang L, Chen W, Petrick NA, Gallas BD. Comparing two correlated C indices with right-censored survival
7 outcome: a one-shot nonparametric approach. *Statistics in medicine* 2015;**34**:685-703.
- 8 50. Rubin DB. Multiple imputation for nonresponse in surveys. Hoboken. In: NJ: John Wiley & Sons; 1987.
- 9 51. Li K-H, Meng X-L, Raghunathan TE, Rubin DB. Significance levels from repeated p-values with multiply-
10 imputed data. *Statistica Sinica* 1991:65-92.
- 11 52. An Y, Tang K, Wang J. Time-Aware Multi-Type Data Fusion Representation Learning Framework for
12 Risk Prediction of Cardiovascular Diseases. *IEEE/ACM Trans Comput Biol Bioinform* 2021;Pp.
- 13 53. Sun L, Pennells L, Kaptoge S, *et al.* Polygenic risk scores in cardiovascular risk prediction: A cohort study
14 and modelling analyses. *PLoS Med* 2021;**18**:e1003498.
- 15 54. Al'Aref SJ, Anchouche K, Singh G, *et al.* Clinical applications of machine learning in cardiovascular
16 disease and its relevance to cardiac imaging. *Eur Heart J* 2019;**40**:1975-1986.
- 17 55. Li Y, Sperrin M, Ashcroft DM, van Staa TP. Consistency of variety of machine learning and statistical
18 models in predicting clinical risks of individual patients: longitudinal cohort study using cardiovascular disease as
19 exemplar. *Bmj* 2020;**371**:m3919.
- 20 56. Gansevoort RT, Correa-Rotter R, Hemmelgarn BR, *et al.* Chronic kidney disease and cardiovascular risk:
21 epidemiology, mechanisms, and prevention. *Lancet* 2013;**382**:339-352.
- 22 57. Emergency Risk Factor Collaboration. C-reactive protein, fibrinogen, and cardiovascular disease
23 prediction. *New England Journal of Medicine* 2012;**367**:1310-1320.
- 24 58. Cook NR. Methods for evaluating novel biomarkers - a new paradigm. *Int J Clin Pract* 2010;**64**:1723-
25 1727.
- 26 59. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and
27 Updating. 1. Aufl. ed. New York, NY: Springer-Verlag; 2009.

- 1 60. Collins R, Reith C, Emberson J, *et al.* Interpretation of the evidence for the efficacy and safety of statin
 2 therapy. *The Lancet* 2016;**388**:2532-2561.
- 3 61. Barbieri S, Mehta S, Wu B, *et al.* Predicting cardiovascular risk from national administrative databases
 4 using a combined survival analysis and deep learning approach. *International journal of epidemiology* 2021;**51**:931-
 5 944.
- 6

7 **Figure legends**

8 **Figure 1 The study design and categories of predictors**

9 (a): The cohort design of the study; (b): Predictors of seven pathways included in different approaches.

10 **Figure 2 The difference of C statistics compared with refitted China-PAR model**

11 The results were given based on the validation set of 31,544.

12 **Figure 3 Calibration plots of difference models by sex^a**

13 The results were given based on the validation set of 31,544.

14 **Figure 4 Distribution of predicted risk given by the XGBoost model and refitted China-PAR model in the validation set**

15

16

17

18

Table 1 Characteristics^a of the study population

| | Overall (N = 215,744) | Men (n = 100,078) | Women (n = 115,666) |
|---------------------------|-----------------------|-------------------|---------------------|
| Demography | | | |
| Age, y | 56.70 (9.59) | 57.10 (9.75) | 56.35 (9.44) |
| Rural | 65,086 (30.34%) | 30,016 (30.15%) | 35,070 (30.51%) |
| Smokers (Current or ever) | 57,961 (26.87%) | 53,861 (53.82%) | 4,100 (3.54%) |
| Finished High school | 108,120 (50.11%) | 55,576 (55.53%) | 52,544 (45.43%) |
| Family history of ASCVD | 1,318 (0.61%) | 701 (0.70%) | 617 (0.53%) |
| Blood pressure | | | |
| SBP, mmHg | 134.45 (16.64) | 134.58 (16.37) | 134.32 (16.88) |
| DBP, mmHg | 82.63 (9.87) | 83.10 (9.90) | 82.18 (9.81) |
| Obesity | | | |
| Waist circumference, cm | 81.76 (7.94) | 83.93 (7.61) | 79.90 (7.73) |
| BMI, kg/m ² | 23.31 (2.87) | 23.44 (2.71) | 23.21 (3.01) |
| Lipid metabolism | | | |

| | | | |
|--|------------------|------------------|------------------|
| Total cholesterol, mmol/L | 4.90 (0.98) | 4.77 (0.97) | 5.01 (0.98) |
| HDL-C, mmol/L | 1.30 (0.34) | 1.25 (0.34) | 1.35 (0.33) |
| TG, mmol/L | 1.61 (1.09) | 1.66 (1.20) | 1.56 (0.99) |
| LDL-C, mmol/L | 2.84 (0.82) | 2.77 (0.81) | 2.90 (0.83) |
| Apo A, mmol/L | 1.22 (0.27) | 1.18 (0.27) | 1.26 (0.27) |
| Apo B, mmol/L | 0.95 (0.25) | 0.95 (0.25) | 0.95 (0.25) |
| Lp(a), mmol/L | 4.87 (0.14) | 4.60 (0.14) | 5.12 (0.15) |
| Glucose metabolism | | | |
| FBG, mmol/L | 5.67 (1.57) | 5.76 (1.72) | 5.60 (1.44) |
| HbA1c, % | 6.86 (1.90) | 6.99 (1.98) | 6.73 (1.82) |
| Diabetes mellitus | 26,090 (12.09%) | 12,364 (12.35%) | 13,726 (11.87%) |
| Renal function | | | |
| eGFR, mL/min/1.73m ² | 98.92 (15.30) | 97.71 (15.28) | 99.94 (15.25) |
| ACR, mg/g | 15.90 (45.36) | 16.32 (48.91) | 15.57 (42.39) |
| Medication | | | |
| Anti-hypertension treatment | 75,857 (35.16%) | 35,590 (35.56%) | 40,267 (34.81%) |
| Anti-hyperlipidemia treatment | 35,561 (16.48%) | 15,662 (15.65%) | 19,899 (17.20%) |
| Anti-hyperglycemia treatment | 22,847 (10.59%) | 10,881 (10.87%) | 11,966 (10.35%) |
| Aspirin treatment | 19,064 (8.84%) | 9,100 (9.09%) | 9,964 (8.61%) |
| Outcome | | | |
| ASVCD events | 6,081 (2.82%) | 3,272 (3.27%) | 2,809 (2.43%) |
| Average follow-up time, years | 5.41 (1.36) | 5.41 (1.51) | 5.41 (1.22) |
| Incidence rate of ASCVD, per million person-years (95% CI) | 6178 (6177-6179) | 7242 (7241-7243) | 5245 (5244-5246) |

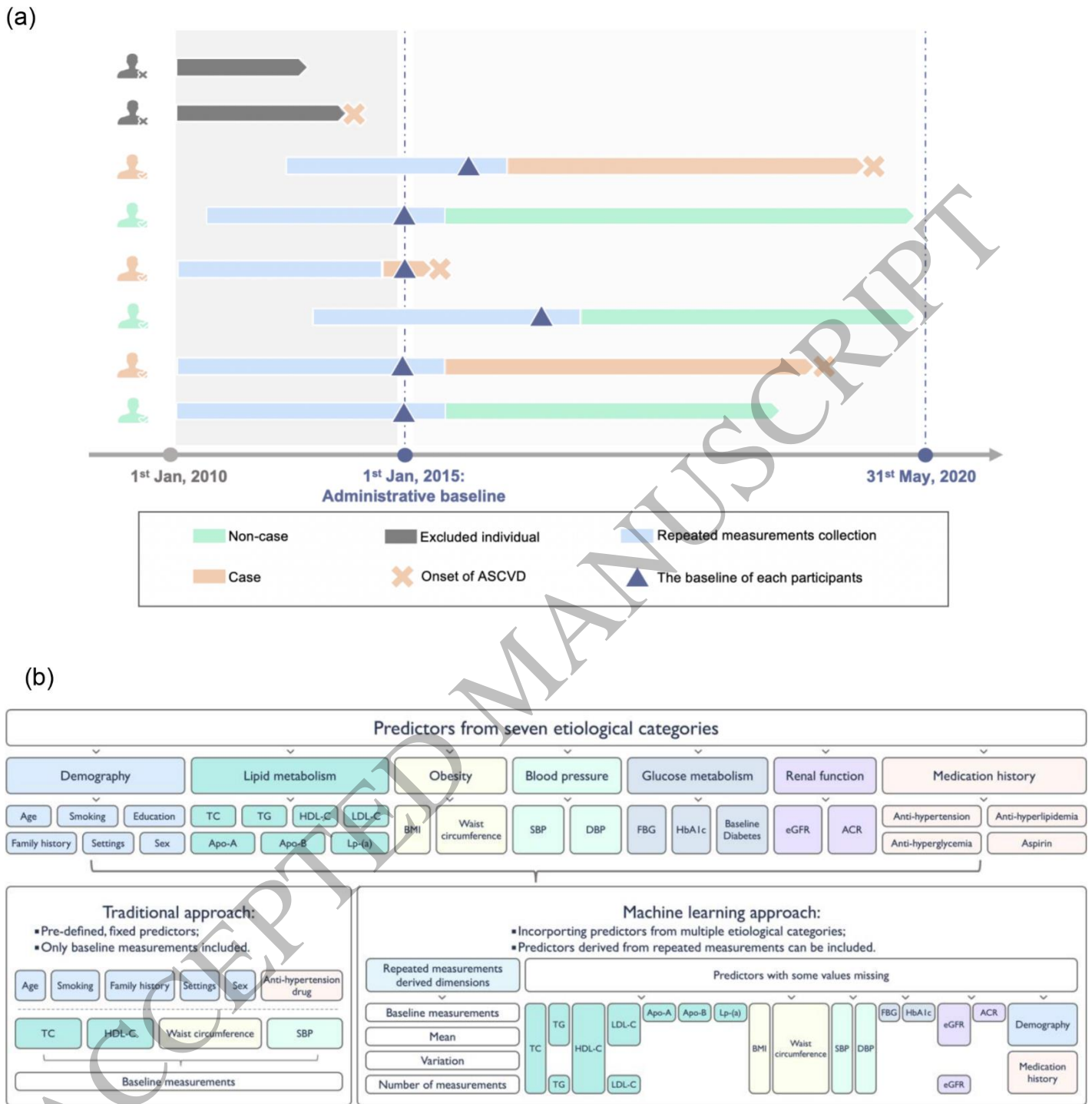
^a Categorical variables were presented by counts and percentages; Continuous variables were presented by means and standard deviations. All the summarized statistics were given based on the complete sets of each predictor.

Table 2 Reclassification of the machine models against refitted China-PAR model^a

| | | XGBoost | | | | NRI (95% CI) | IDI (95% CI) |
|-----------------|---------------------|-----------------|------------------|----------------|--------------|-------------------------|-------------------------|
| | | <2.5% | 2.5%-4.9% | >=5% | Total | | |
| | Refitted PAR | | | | | | |
| Non-case | <2.5% | 14,142 | 270 | 74 | 14,486 | 0.0386 | 0.0174 |
| | 2.5%-4.9% | 2,119 | 2,506 | 323 | 4,948 | (0.0135, 0.0638) | (0.0135, 0.0212) |
| | >=5% | 53 | 1,183 | 3,577 | 4,813 | | |
| | Total | 16,314 | 3,959 | 3,974 | 24,247 | | |
| Case | <2.5% | 185 | 11 | 8 | 204 | | |
| | 2.5%-4.9% | 46 | 114 | 20 | 180 | | |
| | >=5% | 1 | 62 | 522 | 585 | | |
| | Total | 232 | 187 | 550 | 969 | | |
| | | LASSO | | | | | |
| | | <2.5% | 2.5%-4.9% | >=5% | Total | | |
| | Refitted PAR | | | | | | |
| Non-case | <2.5% | 14,147 | 324 | 15 | 14,486 | 0.0278 | 0.0106 |
| | 2.5%-4.9% | 1,057 | 3,543 | 348 | 4,948 | (0.0066, 0.0489) | (0.0081, 0.0131) |
| | >=5% | 3 | 826 | 3,984 | 4,813 | | |
| | Total | 15,207 | 4,693 | 4,347 | 24,247 | | |
| Case | <2.5% | 188 | 14 | 2 | 204 | | |
| | 2.5%-4.9% | 22 | 132 | 26 | 180 | | |
| | >=5% | 0 | 41 | 544 | 585 | | |
| | Total | 210 | 187 | 572 | 969 | | |

^aThe results were given based on the subjects who were not censored (25,216) from the validation set of 31,544.

Figure 1 The study design and categories of predictors



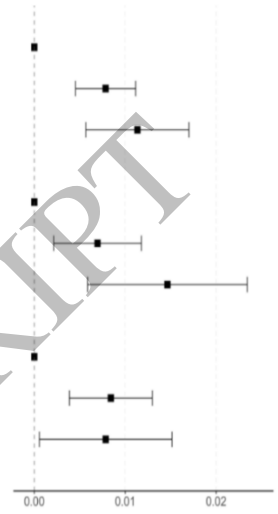
1 (a): The cohort design of the study; (b): Predictors of seven pathways included in different approaches.

2
3
4

Figure 1
188x205 mm (x DPI)

Figure 2 The difference of C statistics compared with refitted China-PAR model^a

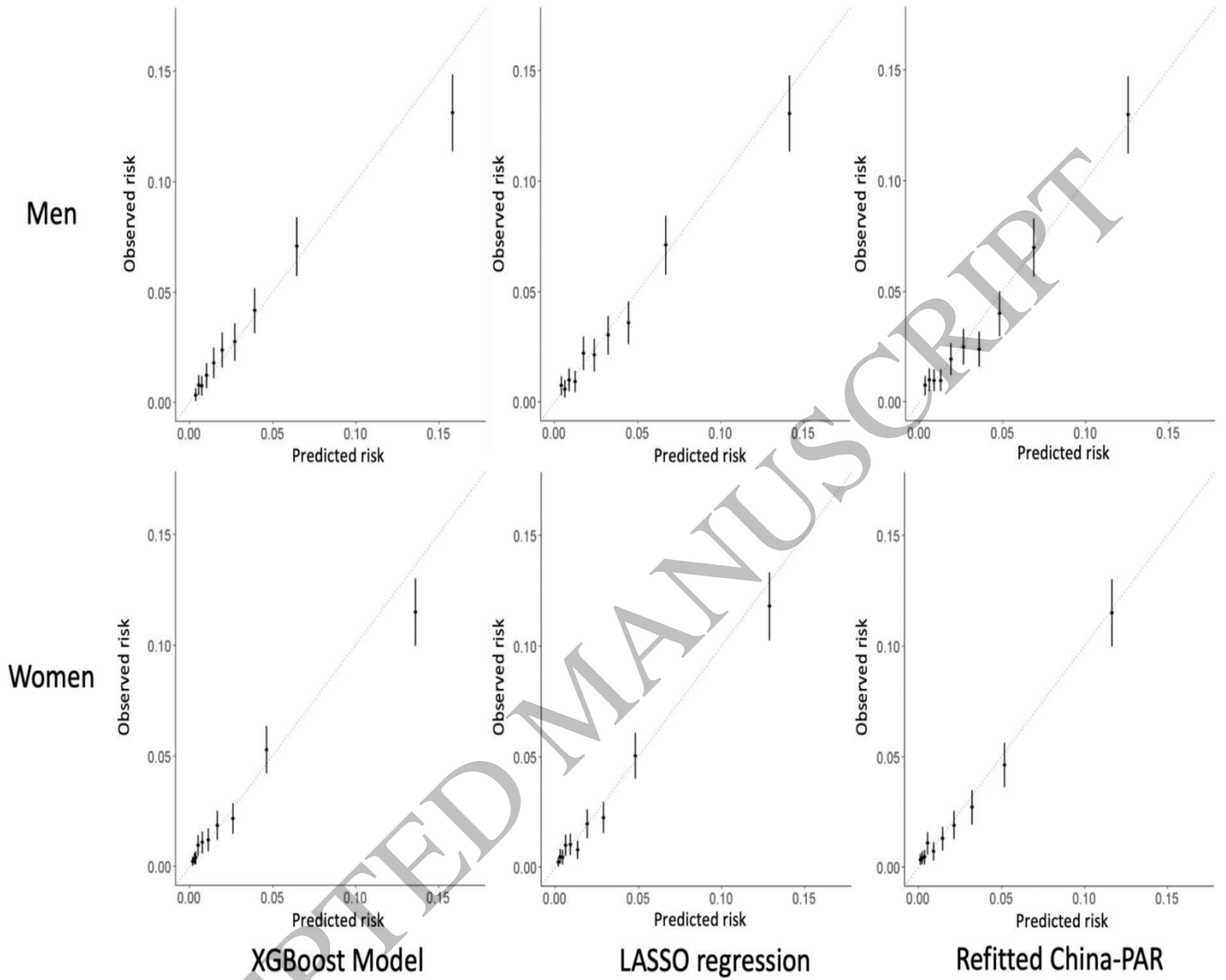
| Sex | Model | C statistics (95% CI) | Difference in C statistics | P |
|---------|--------------------------|-------------------------|----------------------------|---------|
| Overall | Refitted China-PAR model | 0.7805 (0.7657, 0.7953) | Reference | |
| | LASSO regression | 0.7883 (0.7737, 0.8029) | 0.00784 (0.00453, 0.01115) | <0.0001 |
| | XGBoost model | 0.7918 (0.7776, 0.8060) | 0.01134 (0.00567, 0.01700) | <0.0001 |
| Men | Refitted China-PAR model | 0.7554 (0.7340, 0.7767) | Reference | |
| | LASSO regression | 0.7623 (0.7415, 0.7831) | 0.00695 (0.00214, 0.01177) | 0.0047 |
| | XGBoost model | 0.7700 (0.7502, 0.7898) | 0.01464 (0.00586, 0.02342) | 0.0011 |
| Women | Refitted China-PAR model | 0.7992 (0.7780, 0.8205) | Reference | |
| | LASSO regression | 0.8077 (0.7866, 0.8287) | 0.00842 (0.00386, 0.01298) | 0.0003 |
| | XGBoost model | 0.8071 (0.7861, 0.8281) | 0.00785 (0.00056, 0.01514) | 0.0349 |



^aThe results were given based on the validation set of 31,544.

Figure 2
262x113 mm (x DPI)

Figure 3 Calibration plots of difference models by sex^a



^aThe results were given based on the validation set of 31,544.

Figure 3
251x184 mm (x DPI)

Figure 4 Distribution of predicted risk given by the XGBoost model and refitted China-PAR model in the validation set

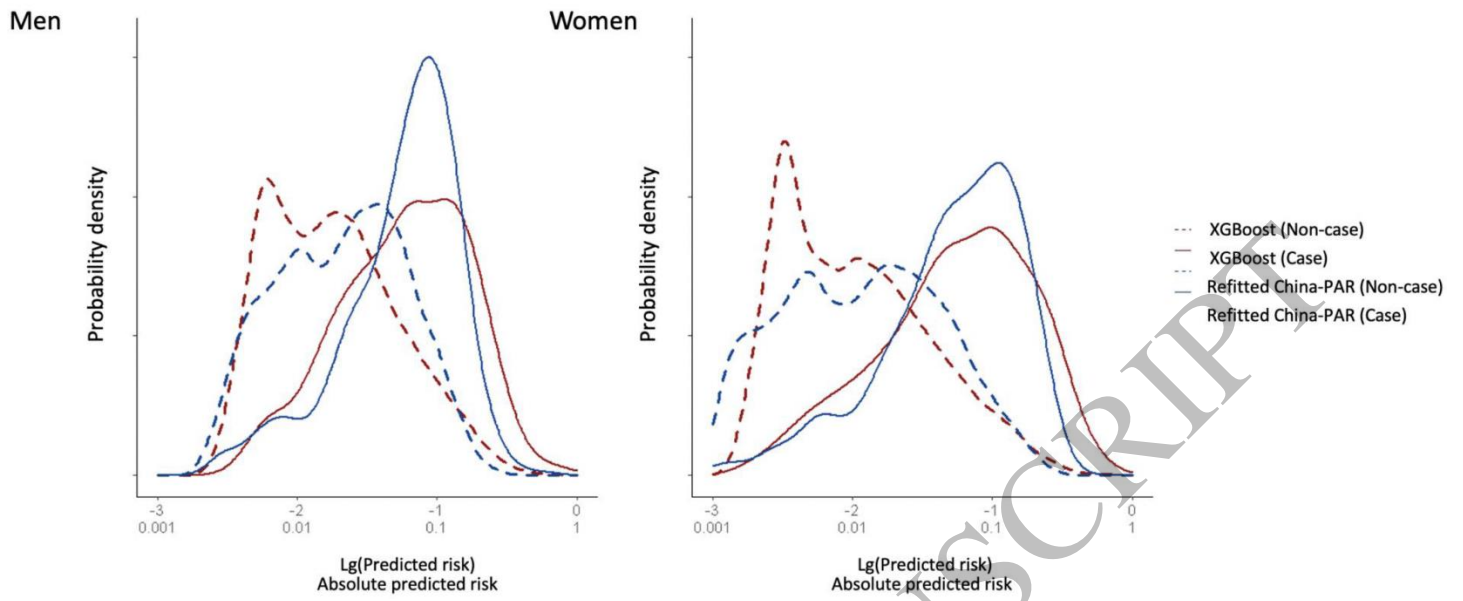


Figure 4

1
2

ACCEPTED MANUSCRIPT

Downloaded from <https://academic.oup.com/ehjdh/advance-article/doi/10.1093/ehjdh/zdad058/7318170> by guest on 01 November 2023

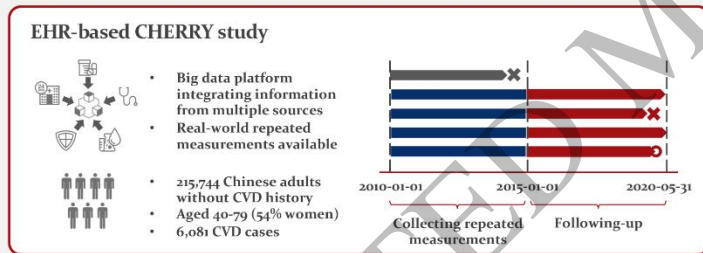
Key Question: Can incorporating irregular, repeated measurements from real-world data into machine learning models improve current cardiovascular risk prediction?

Key Findings: In a cohort with 215,744 Chinese adults, after incorporating 25 repeated-measured cardiovascular risk predictors from electronic health records, the XGBoost model significantly improved the discrimination (C-statistics: 0.792 from 0.781), reclassification (NRI: 3.9%), and calibration compared with the guideline-recommended model.

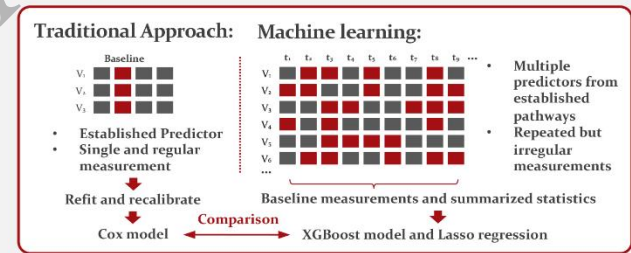
Take-home Message: Machine learning can potentially leverage real-world irregular, repeated measurements to improve current cardiovascular risk prediction.

Can incorporating repeated measurements from real-world data into machine learning models improve current cardiovascular risk prediction?

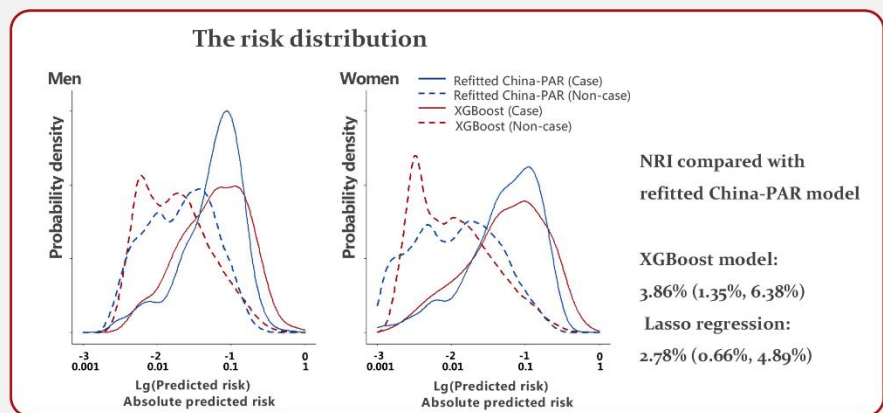
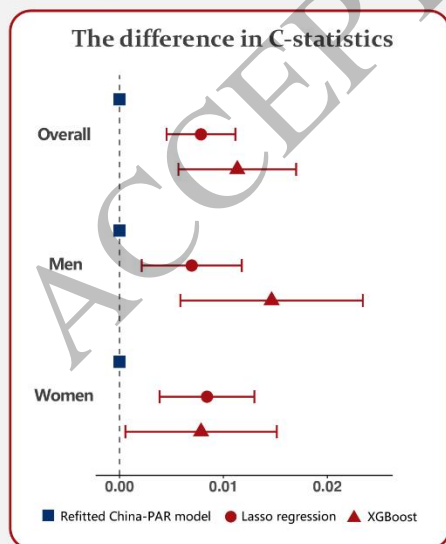
Population and Data:



Design:




Findings:



Conclusion:

Machine learning with repeated real-world data could improve cardiovascular risk prediction on discrimination and reclassification to identify the high-risk population correctly compared with current traditional approach.

Transmissibility quantification of norovirus outbreaks in 2016–2021 in Beijing, China

Yu Wang^{1,2} | Zhiyong Gao²  | Qingbin Lu³  | Baiwei Liu² | Lei Jia² |
Lingyu Shen² | Yi Tian²  | Weihong Li² | Hanqiu Yan² | Daitao Zhang² |
Peng Yang² | Liqun Fang⁴  | Quanyi Wang²  | Fuqiang Cui^{1,3,5} 

¹Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing, China

²Institute for Infectious Disease and Endemic Disease Control, Beijing Center for Disease Prevention and Control, Beijing, China

³Department of Laboratory Science and Technology & Vaccine Research Center, School of Public Health, Peking University, Beijing, China

⁴State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, Beijing, China

⁵Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing, China

Correspondence

Fuqiang Cui, No.38 Xueyuan Rd., Haidian District, Beijing, China.

Email: cui fuq@bjmu.edu.cn

Funding information

Capital's Funds for Health Improvement and Research, Grant/Award Number: 2022-1G-3014; National Key Research and Development Program of China, Grant/Award Number: 2021ZD0114103

Abstract

The transmissibility is a crucial feature for norovirus, yet its quantitative estimation has been limited. Our objective was to estimate the basic reproduction number (R_0) of norovirus and investigate its variation characteristics. Norovirus outbreaks reported from September 2016 to August 2021 in Beijing were analyzed. The susceptible-infected-removed compartment model was established to estimate R_0 . Linear regression models and logistic regression models were used to explore the factors affecting the transmissibility of norovirus. The overall median R_0 of norovirus was estimated as 2.1 (interquartile range [IQR] 1.8–2.5), with 650 norovirus outbreaks. The transmissibility of norovirus varied by year, outbreak setting and genotype. The R_0 of norovirus during September 2019 to August 2020 (median 2.1, IQR 1.8–2.4) and September 2020 to August 2021 (median 2.0, IQR 1.7–2.3) was lower than that of September 2016 to August 2017 (median 2.3, IQR 1.8–2.7) ($\beta = 0.94$, $p = 0.05$; $\beta = 0.93$, $p = 0.008$). The R_0 of norovirus for all other settings was lower than that for kindergarten (median 2.4, IQR 2.0–2.9) (primary school: median 2.0, IQR 1.7–2.4, $\beta = 0.94$, $p = 0.001$; secondary school: median 1.7, IQR 1.5–2.0, $\beta = 0.87$, $p < 0.001$; college: median 1.7, IQR 1.5–1.8, $\beta = 0.89$, $p = 0.03$; other closed settings: median 1.8, IQR 1.5–2.0, $\beta = 0.90$, $p = 0.004$). GII.2[P16] outbreaks had a median R_0 of 2.2 (IQR 1.8–2.7), which was higher than that for GII.6[P7] outbreaks (median 1.8, IQR: 1.8–2.0, odds ratio = 0.19, $p = 0.03$; GII.2[P16] as reference) and mixed-genotype outbreaks (median 1.7, IQR: 1.5–1.8, $\beta = 0.92$, $p = 0.02$; mixed-genotype as reference). In kindergartens and primary schools, norovirus shows increased transmissibility, emphasizing the vulnerable population and high-risk settings. Furthermore, the transmissibility of norovirus may change over time and with virus evolution, necessitating additional research to uncover the underlying mechanisms.

KEYWORDS

basic reproduction number (R_0), epidemiology, genotype, norovirus, outbreaks

1 | INTRODUCTION

Norovirus is a well-known pathogen that causes acute gastroenteritis (AGE) outbreaks and sporadic cases.^{1–3} The United States estimated that norovirus led to 900 deaths, 110 000 hospitalizations, 470 000 emergency department visits, and 2.3 million ambulatory clinic encounters annually,⁴ and 47.0% of AGE outbreaks were caused by norovirus in 2009–2017.² In China, norovirus accounted for a large proportion (89.0%) of AGE outbreaks.³

Transmissibility is an important feature of infectious diseases, and the basic reproduction number (R_0) is widely used to quantify the transmissibility of infectious diseases. R_0 is the average number of secondary cases that were generated by a primary case during the average illness duration in a completely susceptible population. In real-world studies, the estimated R_0 varied, and the data used were collected in different contexts. For example, a previous study reported a median R_0 of 2.75 based on the data of more than 7000 norovirus outbreaks in the United States.² However, a study from China obtained a mean uncontrolled R_0 of 12.2 for norovirus outbreaks with 20 or more epidemiologically linked cases of norovirus infection.⁵ Several studies estimated the R_0 based on the data at the population level, which usually obtained an R_0 lower than that in studies with outbreak data.⁶ One study explored the variation in transmissibility of norovirus and showed that R_0 might vary by outbreak setting and season.²

A norovirus outbreak surveillance network was established in Beijing in 2014, before the release of national guidelines on outbreak

investigation, prevention, and control of norovirus infections.⁷ The transmissibility of norovirus might differ in various genotypes. However, little was known about this aspect in the different genotypes of norovirus. Furthermore, studies on the transmissibility of norovirus have not yet been undertaken in Beijing. Therefore, we aimed to quantify the R_0 of norovirus and further explore whether its transmissibility changes with time, genotypes, or other factors.

2 | MATERIALS AND METHOD

2.1 | Source of data

Data were obtained from the Beijing Center for Disease Control and Prevention (CDC). Once a potential norovirus outbreak was reported, Beijing CDC and district CDCs would conduct epidemiological investigation, specimen collection and etiological detection, and also guide these settings to take measures to control the outbreak.⁸ Detailed information about the data collection procedures of norovirus outbreaks is shown in Figure 1.

2.2 | Inclusion and exclusion criteria of outbreaks

A norovirus outbreak was defined as an outbreak with 10 or more epidemiologically linked AGE cases, and at least two of these cases tested positive for norovirus on laboratory examination. AGE was

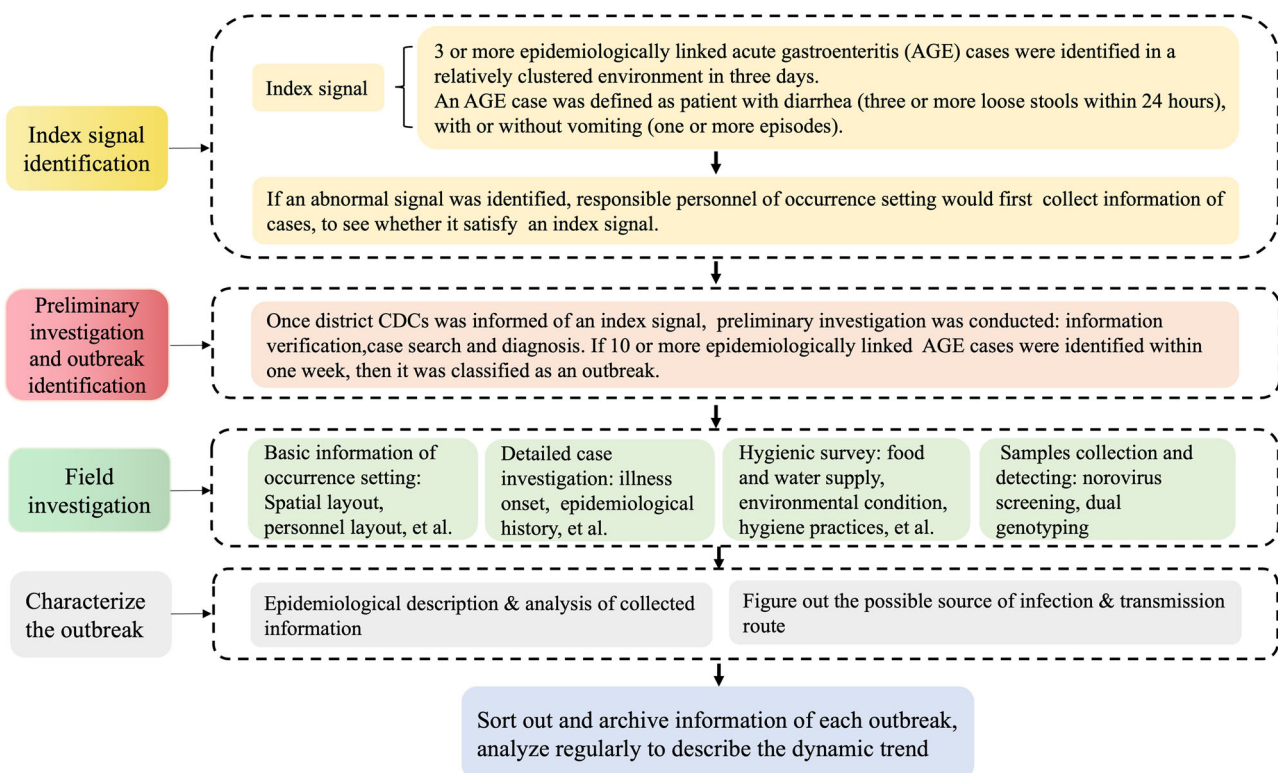


FIGURE 1 Data collection procedures of norovirus outbreaks in Beijing, China.

defined as the development of diarrhea (three or more loose stools within 24 h) with or without vomiting (one or more episodes). The norovirus outbreak data between September 1, 2016 and August 31, 2021 were obtained in our study.

In this analysis, the outbreak data comprised the daily count of new cases, and only the incidence data before the implementation of interventions were used to estimate the R_0 . In the R_0 estimation, we excluded outbreaks with no core information such as onset date, number of susceptible persons, or time series data for analysis.

2.3 | Data collection

The information on norovirus outbreak data collection included onset of outbreak, occurrence region, type of occurrence setting, case number, number of susceptible populations at the start of the outbreak, attack rate, transmission mode, norovirus genogroup and genotype, onset date of each case involved in the outbreak, date when the response was initiated by health department.

A surveillance year was defined as the period from September 1 to August 31 of the following year. The meteorological seasons were divided into four categories: spring (March to May), summer (June to August), autumn (September to November), and winter (December to February). Three occurrence regions were classified into urban, suburban, and outer suburbs according to their distance from the city center and economic development level (Figure 2). Outbreak settings

were referred to as places where infectious sources were introduced and norovirus was transmitted. As most outbreaks occurred in schools, the school setting was subdivided into four categories: kindergartens, primary schools, secondary schools, colleges; and other places included companies or institutions, summer camps or after-school training camps, residential communities, hospitals, events or training groups, nursing homes, child welfare homes, and hotels. The transmission modes were categorized as person-to-person contact, foodborne, waterborne, and unknown. The transmission mode of a norovirus outbreak was initially determined by the staff in district CDCs who were responsible for field epidemiological investigation, and further verified by experts in Beijing CDC. The time period was divided into three phases: “before the coronavirus disease 2019 (COVID-19) outbreak” (September 1, 2016 to December 31, 2019); “extremely strict policy due to COVID-19” (January 1, 2020 to June 30, 2020), during which extremely strict measures were taken, such as city lockdown, schools were closed and implemented online teaching, which resulted in a widespread decline in social activity; and “strict policy due to COVID-19 control” (July 1, 2020 to August 31, 2021), during this stage, social activity gradually began to resume, the measures adopted were more targeted and less extensive, with the minimum impact on the lives of the public. Attack rate was divided into three levels (low, medium, and high) according to its distribution. Timeliness of outbreak response was the time interval between onset date of the first case and the outbreak reporting date, and was classified as early, medium, and late.

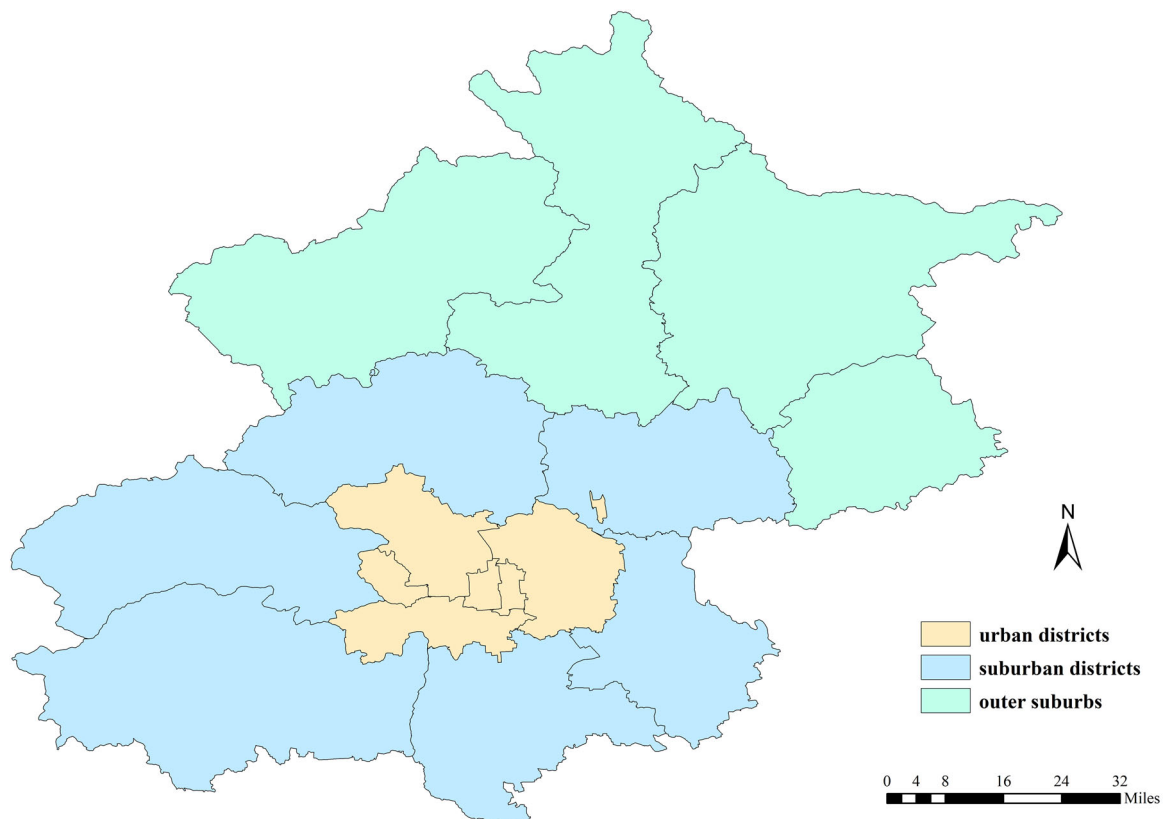


FIGURE 2 The distribution of urban districts, suburban districts, and outer suburbs in Beijing, China.

In addition, the temperature and precipitation data were obtained from the daily observations of the National Oceanic and Atmospheric Administration.⁹ Temperature was regarded as continuous variable, and the precipitation was regarded as categorical variable.

2.4 | R_0 estimation

The R_0 for each outbreak was calculated using the susceptible-infected-removed (SIR) compartment model of infection dynamics. This model could roughly fit the norovirus outbreak epidemic in a relatively closed population by establishing a set of ordinary differential equations and estimating the R_0 based on the outbreak data and maximum likelihood estimates.¹⁰ The established ordinary differential equation was:

$$dS = -\beta \times S \times I,$$

$$dI = \beta \times S \times I - \gamma \times I,$$

$$dR = \gamma \times I,$$

where S was the number of susceptible persons at the start of the outbreak, and it was equal to the size of involved population. I was the number of infected and also infectious persons. R was the number of removed persons who was not infectious and could not be infected in a set of time period. β was the probability that a susceptible person would become infected after contact with an infected person. γ was the probability that the infected person moves out of the compartment. The SIR compartment model assumed that the involved population was closed and homogeneous mixing. The parameters of β and γ were estimated with time series of outbreak data and maximum likelihood estimation.

2.5 | Regression analysis

Categorical variables are presented as counts and percentages, and continuous variables as medians and interquartile range (IQR). Comparisons of differences in continuous dependent variables among groups are conducted using the Kruskal–Wallis H test and Dunn's multiple comparisons test. To explore the factors that affected the norovirus transmissibility, linear regression model was fitted to the log-transformed estimated R_0 , and logistic regression model was fitted to the transmissibility outcome of three levels (high vs. low transmissibility and medium vs. low transmissibility). The candidate factors included year, season (or temperature and precipitation), occurrence region, outbreak setting, transmission mode, and genotype. We selected the genotypes that were responsible for more than 10 norovirus outbreaks and categorized them as GII.2[P16], GII.3[P12], GII.4[P31], GII.6[P7], GII.17[P17], GI.6[P11], mixed-genotype, other GI, and other GII. The same variables were included in the linear regression models and logistic

regression models, and all variables were categorical except temperature. For each variable, the group with the most norovirus outbreaks reported was assigned as the reference group. Before applying the linear regression model, we tested the applicable premise. Weighted least squares were used to produce robust estimates, accounting for heteroscedasticity and nonnormally distributed residuals. Variable selection was performed by comprehensively considering the results of stepwise regression, full subset regression, and professional significance analyses. The model with the lowest Akaike information criterion (AIC) and Bayesian information criterion (BIC) values were selected. Data sorting and analysis were performed using Microsoft Excel and R version 4.2.1. Statistical significance was assumed at $p < 0.05$.

3 | RESULTS

3.1 | Overview of norovirus outbreaks from 2016 to 2021

From September 2016 to August 2021, 738 norovirus outbreaks were reported in Beijing, and a set of R_0 was estimated with 650 ones (88.1%, 650/738). For the 650 norovirus outbreaks, the median outbreak size was 15 cases (IQR 12–20), and the median attack rate was 29.3% (IQR 14.3%–40.4%).

The number of reported norovirus outbreaks varied over the years with the highest in September 2016 to August 2017 (253, 38.9%) (Table 1). Norovirus outbreaks occurred mainly in spring (288, 44.3%) and autumn (184, 28.3%). The majority of the norovirus outbreaks (413, 63.5%) reported in urban districts, which had a higher attack rate (median 31.4%, IQR: 20.0%–40.6%) compared with the other districts (suburban districts: median 23.9%, IQR: 10.4%–38.1%; outer suburbs: median 23.0%, IQR: 8.6%–41.1%), although the outbreak size (median 15, IQR: 12–18) was smaller than other districts (suburban district: median 17, IQR: 13–28; outer suburbs: median 17, IQR: 14–27). Most outbreaks occurred in schools (620, 95.4%), especially in kindergartens (288, 44.3%) and primary schools (264, 40.6%), where outbreaks had higher attack rates (kindergartens: median 37.5%, IQR: 28.0%–47.4%; primary schools: median 25.0%, IQR: 12.5%–34.6%), but smaller scales (kindergartens: median 14 IQR: 11–17; primary schools: median 16 IQR: 13–24) compared with that in secondary schools (attack rate: median 8.4%, IQR: 5.1%–16.2%; scale: median 25, IQR: 17–40) and colleges (attack rate: median 12.4%, IQR: 1.5%–18.0%; scale: median 32, IQR: 23–43). Among the 637 outbreaks with known transmission modes, 94.2% occurred through person-to-person contact, 5.8% were foodborne. Foodborne outbreaks tended to have larger outbreak size (median 25, IQR 19–42) and lower attack rate (median 16.5%, IQR 7.2%–26.8%). With regard to the genogroups of reported norovirus outbreaks, 88.8% (577/650) were caused by GII, 9.4% (61/650) were caused by GI, and 1.8% (12/650) were caused by a combination of GI and GII. Of the 433 norovirus outbreaks whose dual-typing information were successfully obtained, 60.3% (261/433)

TABLE 1 Norovirus outbreaks reported in 2016–2021 in Beijing, China.

| Characteristics | n (%) | Case number | | Attack rate (%) | | R ₀ | |
|--------------------------|------------|--------------|--------|------------------|-----------|----------------|---------|
| | | Median (IQR) | Range | Median (IQR) | Range | Median (IQR) | Range |
| Year | | | | | | | |
| Sep 2016–Aug 2017 | 253 (38.9) | 16 (13–23) | 10–106 | 32.3 (14.2–42.3) | 1.3–79.2 | 2.3 (1.8–2.7) | 1.2–7.0 |
| Sep 2017–Aug 2018 | 71 (10.9) | 14 (11–22) | 10–127 | 27.8 (16.7–36.2) | 2.5–73.3 | 2.1 (1.8–2.5) | 1.3–4.7 |
| Sep 2018–Aug 2019 | 169 (26.0) | 15 (12–21) | 10–155 | 23.9 (12.1–37.5) | 1.6–71.4 | 2.0 (1.7–2.4) | 1.1–5.7 |
| Sep 2019–Aug 2020 | 57 (8.8) | 14 (11–17) | 10–77 | 32.3 (23.8–38.5) | 0.6–72.5 | 2.1 (1.8–2.4) | 1.4–6.0 |
| Sep 2020–Aug 2021 | 100 (15.4) | 15 (12–18) | 10–92 | 28.2 (16.0–41.7) | 0.5–76.5 | 2.0 (1.7–2.3) | 1.4–5.4 |
| Season | | | | | | | |
| Spring | 288 (44.3) | 15 (12–21) | 10–107 | 28.8 (12.5–40.5) | 0.5–79.2 | 2.1 (1.8–2.6) | 1.2–6.0 |
| Summer | 77 (11.9) | 15 (12–25) | 10–61 | 27.7 (13.9–40.4) | 1.4–73.1 | 2.2 (1.7–2.5) | 1.1–7.0 |
| Autumn | 184 (28.3) | 15 (12–18) | 10–155 | 30.9 (16.8–40.1) | 0.6–76.5 | 2.1 (1.8–2.5) | 1.2–6.0 |
| Winter | 101 (15.5) | 15 (12–19) | 10–127 | 28.6 (16.9–40.0) | 3.0–78.6 | 2.1 (1.7–2.5) | 1.2–5.8 |
| Occurrence region | | | | | | | |
| Urban district | 413 (63.5) | 15 (12–18) | 10–145 | 31.4 (20.0–40.6) | 0.5–78.6 | 2.1 (1.8–2.6) | 1.1–6.0 |
| Suburban district | 197 (30.3) | 17 (13–28) | 10–155 | 23.9 (10.4–38.1) | 1.3–79.2 | 2.0 (1.7–2.5) | 1.3–7.0 |
| Outer suburbs | 40 (6.2) | 17 (14–27) | 10–67 | 23.0 (8.6–41.1) | 3.0–57.6 | 2.1 (1.6–2.4) | 1.2–5.8 |
| Outbreak setting | | | | | | | |
| Kindergarten | 288 (44.3) | 14 (11–17) | 10–48 | 37.5 (28.0–47.4) | 6.1–79.2 | 2.4 (2.0–2.9) | 1.4–7.0 |
| Primary school | 264 (40.6) | 16 (13–24) | 10–155 | 25.0 (12.5–34.6) | 1.5–72.5 | 2.0 (1.7–2.4) | 1.2–5.0 |
| Secondary school | 56 (8.6) | 25 (17–40) | 10–92 | 8.4 (5.1–16.2) | 0.5–42.3 | 1.7 (1.5–2.0) | 1.2–3.0 |
| College | 12 (1.9) | 32 (23–43) | 10–61 | 12.4 (1.5–18.0) | 0.6–29.0 | 1.7 (1.5–1.8) | 1.4–2.0 |
| Other closed settings | 30 (4.6) | 16 (14–25) | 10–107 | 14.1 (8.7–25.0) | 1.4–55.0 | 1.8 (1.5–2.0) | 1.1–3.4 |
| Transmission mode | | | | | | | |
| Person-to-person contact | 600 (92.3) | 15 (12–19) | 10–155 | 29.8 (15.0–40.7) | 0.5–79.2 | 2.1 (1.8–2.6) | 1.1–7.0 |
| Foodborne | 37 (5.7) | 25 (19–42) | 11–107 | 16.5 (7.2–26.8) | 1.4–72.5 | 1.9 (1.7–2.1) | 1.4–3.2 |
| Unknown | 13 (2.0) | 14 (11–23) | 10–31 | 33.3 (21.2–35.9) | 4.6–78.6 | 2.3 (1.9–3.0) | 1.4–3.8 |
| Genogroup | | | | | | | |
| GI | 61 (9.4) | 16 (13–24) | 10–145 | 22.6 (11.6–29.8) | 0.6–47.2 | 1.9 (1.7–2.1) | 1.4–5.7 |
| GII | 577 (88.8) | 15 (12–19) | 10–155 | 30.9 (15.3–41.7) | 0.5–79.2 | 2.1 (1.8–2.6) | 1.1–7.0 |
| GI and GII | 12 (1.8) | 17 (12–21) | 10–80 | 12.0 (7.7–17.1) | 3.0–39.3 | 1.6 (1.5–1.7) | 1.4–3.4 |
| Genotype | | | | | | | |
| GII.2[P16] | 261 (60.3) | 15 (12–19) | 10–155 | 32.5 (16.1–43.2) | 1.3–79.2 | 2.2 (1.8–2.7) | 1.4–6.0 |
| GII.3[P12] | 22 (5.0) | 14 (13–17) | 10–30 | 42.7 (21.5–49.5) | 13.0–76.5 | 2.3 (1.9–2.6) | 1.6–3.2 |
| GI.6[P11] | 19 (4.4) | 17 (15–26) | 11–145 | 19.5 (13.3–26.7) | 3.3–34.3 | 1.9 (1.8–2.1) | 1.4–3.6 |
| GII.17[P17] | 19 (4.4) | 20 (12–28) | 10–77 | 14.3 (8.4–30.0) | 1.4–51.7 | 1.9 (1.5–2.2) | 1.3–3.8 |
| GII.6[P7] | 16 (3.7) | 13 (11–26) | 10–55 | 28.6 (15.5–35.4) | 8.3–50.9 | 1.8 (1.8–2.0) | 1.4–3.6 |
| GII.4[P31] | 13 (3.0) | 14 (11–16) | 10–27 | 29.0 (21.4–32.3) | 9.2–44.1 | 2.1 (2.0–2.6) | 1.7–3.2 |
| Mixed genotype | 29 (6.7) | 17 (12–24) | 10–127 | 16.4 (9.8–33.3) | 1.6–59.3 | 1.7 (1.5–1.8) | 1.3–3.6 |
| Other genotypes | 54 (12.5) | 16 (13–25) | 10–92 | 24.5 (11.1–33.1) | 0.6–64.9 | 2.0 (1.7–2.4) | 1.3–5.7 |

(Continues)

TABLE 1 (Continued)

| Characteristics | n (%) | Case number | | Attack rate (%) | | R_0 | |
|--|------------|--------------|--------|------------------|-----------|---------------|---------|
| | | Median (IQR) | Range | Median (IQR) | Range | Median (IQR) | Range |
| Time period | | | | | | | |
| Before COVID-19 (Sep 2016–Dec 2019) | 547 (84.1) | 15 (12–21) | 10–155 | 29.4 (14.2–40.4) | 0.6–79.2 | 2.1 (1.8–2.6) | 1.1–7.0 |
| Extremely strict policy due to COVID-19 (Jan–Jun 2020) | 3 (0.5) | 14 (12–23) | 10–31 | 33.3 (25.1–33.3) | 16.8–33.3 | 1.9 (1.8–2.0) | 1.7–2.1 |
| Strict policy due to COVID-19 (Jul 2020–Aug 2021) | 100 (15.4) | 15 (12–18) | 10–92 | 28.2 (16.0–41.7) | 0.5–76.5 | 2.0 (1.7–2.3) | 1.4–5.4 |
| Attack rate | | | | | | | |
| Low ($\leq 30\%$) | 336 (51.7) | 17 (13–29) | 10–155 | 15.0 (9.1–23.9) | 0.5–29.8 | 1.8 (1.6–2.1) | 1.1–3.4 |
| Medium (30%–50%) | 256 (39.4) | 13 (12–16) | 10–38 | 38.7 (34.3–44.3) | 30.0–50.0 | 2.4 (2.1–2.8) | 1.6–5.8 |
| High ($> 50\%$) | 58 (8.9) | 17 (14–19) | 10–46 | 56.8 (53.9–62.0) | 50.9–79.2 | 3.2 (2.7–3.6) | 1.6–7.0 |
| Response timeliness | | | | | | | |
| Early (≤ 1 day) | 244 (37.5) | 15 (12–19) | 10–127 | 30.8 (14.9–42.2) | 0.5–79.2 | 2.3 (1.9–2.8) | 1.3–6.0 |
| Medium (1–3 days) | 297 (45.7) | 15 (12–21) | 10–155 | 28.9 (14.3–39.0) | 1.3–76.5 | 2.1 (1.8–2.5) | 1.1–7.0 |
| Late (> 3 days) | 109 (16.8) | 15 (12–20) | 10–107 | 28.6 (14.2–39.3) | 0.6–78.6 | 1.9 (1.6–2.3) | 1.2–3.8 |

Abbreviation: IQR, interquartile range.

were caused by GII.2[P16]. Outbreaks size of GII.17[P17] (median 20, IQR: 12–28) was higher than that of other six genotypes in Table 1, but its attack rate was lower (median 14.3%, IQR: 8.4%–30.0%).

When extremely tight policy against COVID-19 (offline classroom teaching was suspended) was enacted in the first half of 2020, only five norovirus outbreaks were reported, which decreased by 96.0% compared to the average number of reported norovirus outbreaks during the same period in the previous 3 years (125). When more targeted and less extensive measures were taken since July 2020, the reporting of norovirus outbreaks gradually rebounded (Table 1). However, only a slight difference was observed in the final outbreak size and attack rate among the three phases. The classification of attack rate was defined as low ($\leq 30\%$), medium (30%–50%), and high ($> 50\%$). Low attack rate outbreaks tended to have large scales. Response timeliness was classified as early (≤ 1 day), medium (1–3 days), and late (> 3 days).

3.2 | The estimated R_0 epidemiological features and univariable analysis

The overall median R_0 was 2.1 (IQR: 1.8–2.5), and the range was 1.1–7.0. The estimated R_0 values for outbreaks of different characteristics are shown in Table 1. The scatter distribution of R_0 and the comparison of differences in univariable analysis was shown in Figure 3.

With regard to the year, the median R_0 for norovirus outbreaks reported from September 2016 to August 2017 (median 2.3, IQR: 1.8–2.7) was slightly higher than that of September 2018 to August 2019 (median 2.0, IQR: 1.7–2.4) and September 2020 to August 2021 (median 2.0, IQR: 1.7–2.3) ($p < 0.001$ and $p = 0.03$, respectively). The R_0 for norovirus outbreaks in suburban districts (median 2.0, IQR: 1.7–2.5) was lower than that in urban districts (median 2.1, IQR: 1.8–2.6) ($p = 0.005$). In terms of the outbreak setting, the R_0 for kindergartens (median 2.4, IQR: 2.0–2.9) was higher than that for other kinds of schools (primary school: median 2.0, IQR: 1.7–2.4; secondary school: median 1.7, IQR: 1.5–2.0; college: median 1.7, IQR: 1.5–1.8) and closed settings (median 1.8, IQR: 1.5–2.0) ($p < 0.001$). For transmission mode, R_0 of norovirus outbreaks with person-to-person contact mode (median 2.1, IQR: 1.8–2.6) was higher than that of foodborne ones (median 1.9, IQR: 1.7–2.1) ($p = 0.04$). For the genogroup, the GII norovirus outbreaks had higher R_0 (median 2.1, IQR: 1.8–2.6) than that of GI norovirus outbreaks (median 1.9, IQR: 1.7–2.1) ($p = 0.003$), and also GI and GII outbreaks (median 1.6, IQR: 1.5–1.7) ($p < 0.001$). The R_0 for GII.2[P16] outbreak (median 2.2, IQR: 1.8–2.7) was higher than that of GII.17[P17] outbreak (median 1.9, IQR: 1.5–2.2) ($p = 0.04$) and mixed-genotype ones (median 1.7, IQR: 1.5–1.8) ($p < 0.001$). R_0 distribution was different among groups of attack rate ($p < 0.001$). R_0 of low attack rate outbreaks (median 1.8, IQR: 1.6–2.1) was lower than that of medium attack rate ones (median 2.4, IQR: 2.1–2.8; $p < 0.001$) and high attack rate ones (median 3.2, IQR: 2.7–3.6; $p < 0.001$). Response timeliness of outbreaks was also related with R_0 distribution ($p < 0.001$). R_0 of

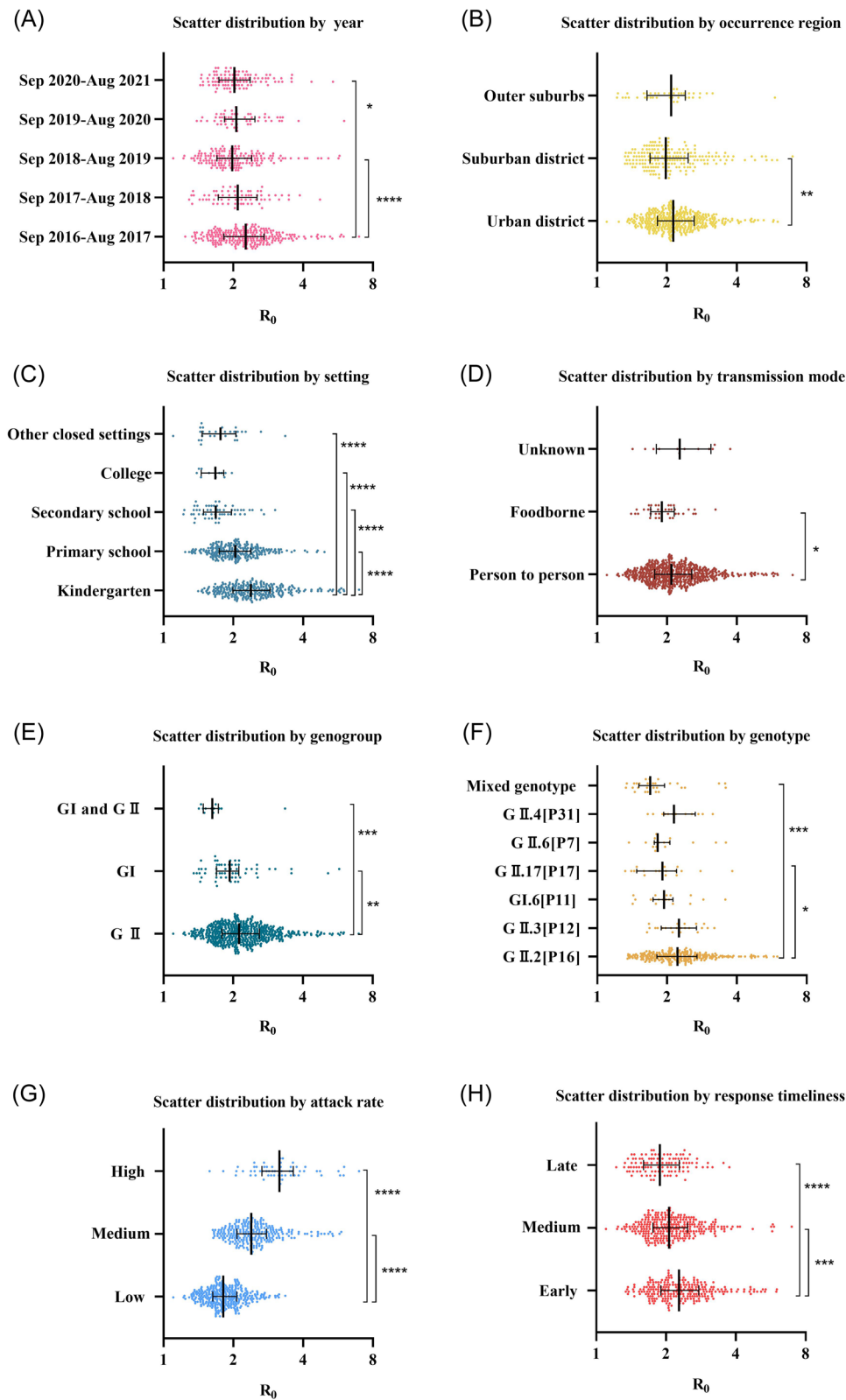


FIGURE 3 Scatter distribution of the estimated R_0 (with median and interquartile range) and comparison of R_0 between groups by characteristics of (A) year, (B) occurrence region, (C) outbreak setting, (D) transmission mode, (E) genogroup, (F) genotype, (G) attack rate, (H) response timeliness. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$.

outbreaks with early response (median 2.3, IQR: 1.9–2.8) was higher than that of medium response outbreaks (median 2.1, IQR: 1.8–2.5; $p < 0.001$) and late response ones (median 1.9, IQR: 1.6–2.3; $p < 0.001$). No statistical difference was found in the R_0 among the groups in terms of season and time period.

3.3 | Multivariable analysis by linear regression and logistic regression models

R_0 was classified as three levels: low ($R_0 < 2.0$), medium ($2.0 \leq R_0 < 2.6$), and high ($R_0 \geq 2.6$), corresponding to the low, medium, and high transmissibility in logistic regression analysis. Low transmissibility group was the control group.

Several factors (year, occurrence region, outbreak setting, genogroup, genotype, transmission mode, attack rate, response timeliness, temperature, precipitation) were included in the linear regression model; among them, year, occurrence region, outbreak setting, genotype, attack rate and response timeliness were selected in the final model (AIC = -1279.6; BIC = -1172.2) (Table 2).

The R_0 for norovirus outbreaks reported from September 2019 to August 2020 (median 2.1, IQR: 1.8–2.4; $\beta = 0.94$, $p = 0.05$) and September 2020 to August 2021 (median 2.0, IQR: 1.7–2.3; $\beta = 0.93$, $p = 0.008$) was lower than that of September 2016 to August 2017 (median 2.3, IQR: 1.8–2.7). No statistical significance was observed in the R_0 for that of September 2017 to August 2018 (median 2.1, IQR: 1.8–2.5) and September 2018 to December 2019 (median 2.0, IQR: 1.7–2.4), compared with that for the September 2016 to August 2017; these findings were also supported by the results of logistic regression model. The R_0 for norovirus outbreaks that occurred in outer suburbs (median 2.1, IQR: 1.6–2.4) was lower than that in urban districts (median 2.1, IQR: 1.8–2.6; $\beta = 0.93$, $p = 0.01$) in linear regression analysis, but the logistic regression model showed no statistical significance of R_0 for outbreaks occurred in suburban districts and outer suburbs compared with that of urban district. The variations in R_0 by outbreak setting was obvious, with a lower R_0 of norovirus in all other settings (primary school: median 2.0, IQR: 1.7–2.4; secondary school: median 1.7, IQR: 1.5–2.0; college: median 1.7, IQR: 1.5–1.8; other closed settings: median 1.8, IQR: 1.5–2.0) compared with that in kindergartens (median 2.4, IQR: 2.0–2.9) (primary school: $\beta = 0.94$, $p = 0.001$; secondary school: $\beta = 0.87$, $p < 0.001$; college: $\beta = 0.89$, $p = 0.03$; other closed settings: $\beta = 0.90$, $p = 0.004$).

For outbreak genotype, mixed genotype outbreaks had a median R_0 of 1.7 (IQR: 1.5–1.8), which was lower than that of GII.2[P16] outbreaks (median 2.2, IQR: 1.8–2.7) ($\beta = 0.92$, $p = 0.02$), but no statistical significance was found for other groups (GII.3[P12]: median 2.3, IQR: 1.9–2.6; GI.6[P11]: median 1.9, IQR: 1.8–2.1; GII.17[P17]: median 1.9, IQR: 1.5–2.2; GII.6[P7]: median 1.8, IQR: 1.8–2.0; GII.4[P31]: median 2.1, IQR: 2.0–2.6) in linear regression analysis. In the logistic regression model, GII.6[P7] outbreaks also had a lower R_0 (median 1.8, IQR: 1.8–2.0) compared with that of GII.2[P16] outbreaks (medium vs. low transmissibility: odds ratio [OR] = 0.19,

$p = 0.03$; high vs. low transmissibility: OR = 0.15, $p = 0.05$). Compared with low attack rate outbreaks (median 1.8, IQR: 1.6–2.1), R_0 of medium attack rate outbreaks (median 2.4, IQR: 2.1–2.8) and high attack rate ones (median 3.2, IQR: 2.7–3.6) was higher ($\beta = 1.27$, $p < 0.001$; $\beta = 1.60$, $p < 0.001$, respectively), which was also supported by the logistic regression model. As for the response timeliness of outbreaks, medium response outbreaks (median: 2.1, IQR: 1.8–2.5) and late response ones (median 1.9, IQR: 1.6–2.3) had lower R_0 compared with the early ones (median 2.3, IQR: 1.9–2.8) ($\beta = 0.93$, $p < 0.001$; $\beta = 0.84$, $p < 0.001$, respectively), and this result was also supported by the logistic regression model.

4 | DISCUSSION

The key epidemiological characteristics of norovirus outbreaks from September 2016 to August 2021 in Beijing were largely unchanged,⁸ predominantly reported in the spring season, affecting kindergartens and primary schools, as well as urban areas, and were transmitted via person-to-person contact. GII remains the primary genogroup responsible for norovirus outbreaks. The emergence of GII.2[P16] norovirus since 2016 resulted in a significant rise in AGE outbreaks in Beijing, and it remains the dominant strain in recent years. The most common setting of norovirus outbreak in Beijing was in schools, which was quite different from that in some countries, where norovirus outbreaks mainly occurred in long-term care facilities or in hospitals. This could be related to some reasons: first, underreporting might exist in hospitals, for outbreaks of nosocomial infections are managed by other health administrative department and are usually not reported to the CDC. Second, there are cultural differences regarding elderly care between China and some other countries. Chinese seniors prefer to be taken care of at home, rather than go to a nursing home. Third, the surveillance work is focused on schools since 2014, and CDCs work with education department to conduct extensive training on outbreak reporting, resulting in improved reporting practices. But this kind of training is not carried out in other kind of settings. Hence, the reporting bias might also exist.

In this study, we conducted quantitative estimation on transmissibility of norovirus, which could add evidence to the understanding of norovirus and guide the prevention and control work. The estimated R_0 was around 2.1, which was similar to the R_0 value calculated by other studies with a data set of many outbreaks.^{2,11} One study used the data of 75 outbreaks that occurred in hospitals and long-term care facilities in England in 2002–2003, and obtained the R_0 values of 2.7 in long-term care facilities and 1.3 in hospitals.¹¹ Another study estimated an R_0 of 2.7, with more than 7000 norovirus outbreaks reported from 2009 to 2017 in various settings (mainly long-term care facilities) in the United States.² The estimated R_0 of this study was lower than that of some studies which concluded from a single norovirus outbreak.^{12,13} For example, one study applied the data of a norovirus outbreak involving 360 cases, and got an R_0 of 8.32.¹³ Such high-estimated R_0 was mainly related to the outbreak data applied, which usually involved a large number of cases and

TABLE 2 Factors affecting the transmissibility of norovirus through establishing linear regression model and logistic regression model.

| Characteristics | Estimated log-linear change in R_0 (95% CI) | <i>p</i> Value | OR (95% CI) of medium VS. low R_0 gr | <i>p</i> Value | OR of $R_0 \geq 2.6$ (95% CI) | <i>p</i> Value |
|--------------------------------------|---|----------------|--|----------------|-------------------------------|----------------|
| Intercept | 2.22 (2.11–2.34) | <0.001 | 1.13 (0.61–2.09) | 0.69 | 0.63 (0.28–1.40) | 0.26 |
| Year | | | | | | |
| Sep 2016–Aug 2017 | Referent | | Referent | | Referent | |
| Sep 2017–Aug 2018 | 0.96 (0.91–1.02) | 0.17 | 1.12 (0.55–2.30) | 0.75 | 0.73 (0.27–1.95) | 0.53 |
| Sep 2018–Aug 2019 | 0.97 (0.92–1.01) | 0.12 | 0.80 (0.46–1.41) | 0.45 | 0.56 (0.26–1.18) | 0.13 |
| Sep 2019–Aug 2020 | 0.94 (0.88–1.00) | 0.05 | 0.56 (0.26–1.22) | 0.15 | 0.29 (0.10–0.82) | 0.02 |
| Sep 2020–Aug 2021 | 0.93 (0.89–0.98) | 0.008 | 0.55 (0.28–1.06) | 0.07 | 0.24 (0.10–0.59) | 0.002 |
| Occurrence region^a | | | | | | |
| Urban district | Referent | | NA | NA | NA | NA |
| Suburban district | 0.98 (0.95–1.02) | 0.37 | | | | |
| Outer suburbs | 0.93 (0.87–0.98) | 0.01 | | | | |
| Occurrence setting | | | | | | |
| Kindergarten | Referent | | Referent | | Referent | |
| Primary school | 0.94 (0.90–0.97) | 0.001 | 0.78 (0.49–1.25) | 0.31 | 0.41 (0.22–0.76) | 0.004 |
| Secondary school | 0.87 (0.82–0.92) | < 0.001 | 0.34 (0.14–0.80) | 0.01 | 0.32 (0.08–1.29) | 0.11 |
| College ^b | 0.89 (0.81–0.99) | 0.03 | 0 (0–0) | 0 | 0 (0–0) | 0 |
| Other closed settings | 0.90 (0.84–0.97) | 0.004 | 0.73 (0.27–1.95) | 0.53 | 0.50 (0.09–2.77) | 0.43 |
| Genotype | | | | | | |
| GII.2[P16] | Referent | | Referent | | Referent | |
| GII.3[P12] | 0.94 (0.86–1.03) | 0.20 | 0.92 (0.29–2.93) | 0.89 | 0.64 (0.14–2.93) | 0.56 |
| GII.4[P31] | 1.09 (0.97–1.23) | 0.14 | 1.76 (0.42–7.40) | 0.44 | 3.28 (0.52–20.64) | 0.21 |
| GII.6[P7] | 0.94 (0.85–1.03) | 0.20 | 0.19 (0.04–0.82) | 0.03 | 0.15 (0.02–1.04) | 0.05 |
| GII.17[P17] | 0.99 (0.91–1.08) | 0.82 | 0.54 (0.14–2.04) | 0.36 | 0.49 (0.06–4.07) | 0.51 |
| GI.6[P11] | 1.01 (0.92–1.10) | 0.89 | 0.57 (0.17–1.97) | 0.38 | 0.65 (0.10–4.13) | 0.65 |
| other GI | 1.00 (0.94–1.07) | 0.97 | 0.73 (0.31–1.75) | 0.49 | 0.64 (0.17–2.51) | 0.53 |
| other GII | 0.99 (0.95–1.02) | 0.46 | 1.03 (0.63–1.70) | 0.89 | 1.01 (0.53–1.94) | 0.97 |
| Mixed genotype | 0.92 (0.85–0.99) | 0.02 | 0.24 (0.07–0.80) | 0.02 | 0.39 (0.07–2.04) | 0.26 |
| Attack rate | | | | | | |
| Low ($\leq 30\%$) | Referent | | | | | |
| Medium (30%–50%) | 1.27 (1.22–1.32) | < 0.001 | 5.68 (3.54–9.09) | < 0.001 | 21.97 (11.09–43.53) | < 0.001 |
| High ($> 50\%$) | 1.60 (1.48–1.72) | < 0.001 | 12.65 (2.61–61.27) | 0.002 | 336.94 (65.80–1725.35) | < 0.001 |
| Response timeliness | | | | | | |
| Early (≤ 1 day) | Referent | | | | | |
| Medium (1–3 days) | 0.93 (0.90–0.96) | < 0.001 | 0.68 (0.43–1.07) | 0.10 | 0.32 (0.18–0.58) | < 0.001 |
| Late (> 3 days) | 0.84 (0.81–0.88) | < 0.001 | 0.42 (0.23–0.79) | 0.007 | 0.05 (0.02–0.13) | < 0.001 |

Note: Bold values indicate statistical significance.

Abbreviations: CI, confidence interval; OR, odds ratio.

^aVariable of occurrence region was not selected in the final model in the logistic regression analysis.

^bFor 12 norovirus outbreaks occurred in college which were reported in Beijing during September 2016 and August 2021, R_0 was all lower than 2.

were special in epidemiological characteristic; hence, publication bias might have existed. Another study in China, identified that the R_0 could be as high as 12.2,⁵ which was close to the transmissibility of measles.¹⁴ This could be also related to the outbreak data applied (number size ≥ 20), model structure established (considering the transmissibility of asymptomatic cases), and parameter setting. However, outbreaks with <20 cases were more common in actual norovirus surveillance^{15,16}; and information on asymptomatic cases in most outbreaks were difficult to obtain during field investigations. Therefore, we assume this study might overestimate the transmissibility of norovirus. The available evidence suggests that R_0 based on the norovirus outbreak data was higher than that calculated using the community surveillance data,⁶ and the R_0 of our study is similar to that of the community surveillance data. It might be due to the fact that the data we utilized encompassed numerous outbreaks spanning a 5-year timeframe. The spread of norovirus reflected by this outbreak data set was closer to the situation in community population, compared with just a single outbreak. Meanwhile, the underestimation of R_0 might exist in our study, and it could be due to two reasons. First, the number of susceptible populations at the start of an outbreak was difficult to determine accurately, and it was usually determined by experienced investigators in district CDCs. Normally, people in contact with cases in a confined space are regarded as susceptible. However, some people might be immune to norovirus infections¹⁷; therefore, we overestimated the susceptible population and underestimated the R_0 . Another reason may be that some infection cases were not recognized or involved in the outbreak data set. The case definition for norovirus outbreaks only included symptomatic cases and excluded cases with diarrhea less than three times a day and asymptomatic infections. This led to the misclassification of infected persons as uninfected and underestimation of R_0 . These were part of the limitations of our study; we could optimize it by collecting more information of the susceptible population and increasing the surveillance sensitivity.

Some characteristics were identified to affect the transmissibility of norovirus (year, occurrence region, outbreak setting, genotype, attack rate and response timeliness). The R_0 of norovirus showed a decreasing trend in recent 2 years (September 2019 to August 2020 and September 2020 to August 2021), compared with that of September 2016 to August 2017. This finding could be due to the enhancement of population immunity after infection. But it was merely one explanation, and additional time-dependent factors might affect the fluctuation of R_0 , necessitating further study to explore. Outbreaks in urban districts had a slightly higher transmissibility compared with those in outer suburbs, which might be related to the large population density in urban districts. The variation in R_0 values among the outbreak settings were obvious. Outbreaks in kindergartens had the highest transmissibility, followed by primary schools, whereas the transmissibility in secondary schools, colleges, and other closed settings slightly varied. This variation was due to the different age groups and various characteristics of population involved in these settings. Kids were more active and unconstrained, and had a higher frequency of contact with peers. In addition, children were more

vulnerable to norovirus infection. Therefore, it was understandable that outbreaks occurred in pediatric group had the highest transmissibility. A similar trend was observed in another study: for schools, kindergartens have the highest transmissibility, followed by primary schools.⁵ Hence, priorities should be set for school settings at lower ages. Our study found GII.2[P16] norovirus was more transmissible than GII.6[P7] and the mixed-genotype norovirus, but the mechanism needs to be further studied. Available study on the transmissibility of different norovirus genotypes was very limited. In one study, the R_0 of GII.6[P7] norovirus was slightly higher than GII.2[P16] ones, but there was no statistical difference.⁵ Another study found no evidence of variation in the estimated R_0 value among norovirus genotypes categorized as GI, GII.4, or non-GII.4.² But it did not offer more information about whether certain dual-genotypes were more transmissible. At the beginning of this study, we assumed that R_0 for different genotypes was different, similar to that observed among SARS-CoV-2 variants.¹⁸ However, the variation of transmissibility among several genotypes was not that obvious. Maybe the sample size of these genotype outbreaks observed in this study was not large enough to observe the differences. We need more data and continuous research to explore this. Besides, outbreaks of different attack rate had different transmissibility. High and median attack rate outbreaks had higher R_0 compared with low attack rate ones. This suggests that the estimated R_0 in this study was reliable for it has good consistency with the attack rate of outbreaks. R_0 was also related with response timeliness. The earlier the response was initiated, the higher the R_0 was. This was mainly attributed to the data we used. The estimation of R_0 required outbreak incidence data which was not influenced by any human intervention. However, in real-world situation, health department would take timely measures to curb the spread of norovirus in the vast majority of outbreaks. Hence, only the incidence data before the implementation of interventions was used to estimate R_0 . And this part of data just reflects the upward phase of the epidemiological curve for an outbreak, which might lead to a higher R_0 estimation. This could be a limitation for our study. Although norovirus outbreaks presented seasonal patterns and different transmission modes,^{15,16} the transmissibility did not vary by season (or temperature and precipitation) or transmission mode in this study.

There were several limitations to our study in addition to those mentioned above. First, we only included norovirus outbreaks with more than 10 cases, which might have caused selection bias. Second, in the SIR dynamic model, we assumed homogeneity in susceptibility, contact pattern, and the ability to infect others in a norovirus outbreak. However, in a real scenario, an individual might not be genetically susceptible to norovirus infections,¹⁹ and there might be super-spreaders who can infect more people. This simplified assumption may not accurately reflect a real-world complex situation. Third, the data used to calculate the R_0 of an outbreak were the incidence data of cases before the implementation of intervention measures, which could not represent the natural and complete course of outbreaks, as discussed above. Hence, further studies are needed to improve the accuracy of the surveillance data and further optimize the estimation.

In summary, norovirus outbreaks still occur frequently even in cities with relatively good sanitation conditions like Beijing, indicating the significant transmissibility of norovirus. Particularly in places where susceptible individuals gather, such as kindergartens and primary schools, norovirus spreads more easily and exhibits higher transmissibility, emphasizing the need for early interventions in these high-risk populations and settings. Furthermore, we have noted difference in R_0 in relation to surveillance year and genotype. This indicates that the transmissibility of norovirus may vary over time and with virus evolution. Consequently, additional research and evidence are crucial to reveal the underlying mechanisms, which will contribute scientific evidence to support the deployment and assessment of interventions, as well as the advancement of vaccines and medications.

AUTHOR CONTRIBUTIONS

Fuqiang Cui and Yu Wang designed the study. Yu Wang, Zhiyong Gao, Baiwei Liu, Lei Jia, Weihong Li, Hanqiu Yan, Lingyu Shen, and Yi Tian were responsible for data collection. Qingbin Lu, Daitao Zhang, Peng Yang, and Quanyi Wang offered technical and material support. Yu Wang analyzed the data. Yu Wang and Fuqiang Cui drafted the manuscript. Fuqiang Cui, Zhiyong Gao, Peng Yang, Quanyi Wang, and Liqun Fang revised the final manuscript. All authors approved the final draft of the manuscript. Fuqiang Cui is the guarantor. The corresponding author attests that all listed authors meet the authorship criteria and that no author meeting the criteria have been omitted.

ACKNOWLEDGMENTS

We would like to thank the staff members in district and municipal Centers for Disease Prevention and Control and medical settings in Beijing for performing the field investigations, specimen collection, laboratory detection, and case reporting. This work was supported by the National Key R&D Program of China under Grant number 2021ZD0114103; and the Capital's Funds for Health Improvement and Research under Grant number 2022-1G-3014. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data will be shared on reasonable request to the corresponding author.

ETHICS STATEMENT

This study was approved by the Ethical Committee of Beijing Center for Disease Prevention and Control, and the requirement for informed consent was waived.

ORCID

Zhiyong Gao  <http://orcid.org/0000-0003-4413-4265>

Qingbin Lu  <http://orcid.org/0000-0002-2804-0827>

Yi Tian  <http://orcid.org/0000-0002-0831-6275>

Liqun Fang  <http://orcid.org/0000-0002-4981-1483>

Quanyi Wang  <http://orcid.org/0000-0001-9552-2503>

Fuqiang Cui  <http://orcid.org/0000-0002-9592-4286>

REFERENCES

- Ahmed SM, Hall AJ, Robinson AE, et al. Global prevalence of norovirus in cases of gastroenteritis: a systematic review and meta-analysis. *Lancet Infect Dis*. 2014;14(8):725-730. doi:10.1016/S1473-3099(14)70767-4
- Steele MK, Wikswo ME, Hall AJ, et al. Characterizing norovirus transmission from outbreak data, United States. *Emerging Infect Dis*. 2020;26(8):1818-1825. doi:10.3201/eid2608.191537
- Lian Y, Wu S, Luo L, et al. Epidemiology of norovirus outbreaks reported to the public health emergency event surveillance system, China, 2014-2017. *Viruses*. 2019;11(4):342. doi:10.3390/v11040342
- Burke RM, Mattison CP, Pindyck T, et al. Burden of norovirus in the United States, as estimated based on administrative data: updates for medically attended illness and mortality, 2001-2015. *Clin Infect Dis*. 2021;73(1):e1-e8. doi:10.1093/cid/ciaa438
- Ai J, Zhu Y, Fu J, et al. Study of risk factors for total attack rate and transmission dynamics of norovirus outbreaks, Jiangsu province, China, from 2012 to 2018. *Front Med (Lausanne)*. 2022;8:786096. doi:10.3389/fmed.2021.786096
- Gaythorpe KAM, Trotter CL, Lopman B, Steele M, Conlan AJK. Norovirus transmission dynamics: a modelling review. *Epidemiol Infect*. 2018;146(2):147-158. doi:10.1017/S0950268817002692
- Liao Q, Ran L, Jin M, et al. [Guidelines on outbreak investigation, prevention and control of norovirus infection (2015)]. *Zhonghua Yu Fang Yi Xue Za Zhi*. 2016;50(1):7-16. Chinese. doi:10.3760/cma.j.issn.0253-9624.2016.01.003
- Gao Z, Liu B, Yan H, et al. Norovirus outbreaks in Beijing, China, from 2014 to 2017. *J Infect*. 2019;79(2):159-166. doi:10.1016/j.jinf.2019.05.019
- Daily observational data [Internet]. Asheville: National Oceanic and Atmospheric Administration; Available from: <https://www.noaa.gov/maps/daily/>
- Ottar N. B. *Epidemics: Models and Data using R*. Springer Cham; 2018:137-157. doi:10.1007/978-3-319-97487-3
- O'Dea EB, Pepin KM, Lopman BA, Wilke CO. Fitting outbreak models to data from many small norovirus outbreaks. *Epidemics*. 2014;6:18-29. doi:10.1016/j.epidem.2013.12.002
- Heijne JCM, Teunis P, Morroy G, et al. Enhanced hygiene measures and norovirus transmission during an outbreak. *Emerging Infect Dis*. 2009;15(1):24-30. doi:10.3201/eid1501.080299
- Xu Y, Zhu Y, Lei Z, et al. Investigation and analysis on an outbreak of norovirus infection in a health school in Guangdong Province, China. *Infect Genet Evol*. 2021;96:105135. doi:10.1016/j.meegid.2021.105135
- Guerra FM, Bolotin S, Lim G, et al. The basic reproduction number (R_0) of measles: a systematic review. *Lancet Infect Dis*. 2017;17(12):e420-e428. doi:10.1016/S1473-3099(17)30307-9
- Calderwood LE, Wikswo ME, Mattison CP, et al. Norovirus outbreaks in long-term care facilities in the United States, 2009-2018: a decade of surveillance. *Clin Infect Dis*. 2022;74(1):113-119. doi:10.1093/cid/ciab808
- Jin M, Wu S, Kong X, et al. Norovirus outbreak surveillance, China, 2016-2018. *Emerging Infect Dis*. 2020;26(3):437-445. doi:10.3201/eid2603.191183
- Frenck R, Bernstein DI, Xia M, et al. Predicting susceptibility to norovirus GII.4 by use of a challenge model involving humans. *J Infect Dis*. 2012;206(9):1386-1393. doi:10.1093/infdis/jis514

18. Liu Y, Rocklöv J. The effective reproductive number of the Omicron variant of SARS-CoV-2 is several times relative to Delta. *J Travel Med.* 2022;29(3):taac037. doi:10.1093/jtm/taac037
19. Marionneau S, Ruvoën N, Le Moullac-Vaidye B, et al. Norwalk virus binds to histo-blood group antigens present on gastroduodenal epithelial cells of secretor individuals. *Gastroenterology.* 2002;122(7):1967-1977. doi:10.1053/gast.2002.33661

How to cite this article: Wang Y, Gao Z, Lu Q, et al. Transmissibility quantification of norovirus outbreaks in 2016–2021 in Beijing, China. *J Med Virol.* 2023;95:e29153. doi:10.1002/jmv.29153

Spatiotemporal cluster of mpox in men who have sex with men: A modeling study in 83 countries

Weijing Shang | Guiying Cao | Yu Wu | Liangyu Kang | Yaping Wang |
Peng Gao | Jue Liu  | Min Liu 

School of Public Health, Peking University,
Beijing, China

Correspondence

Min Liu, School of Public Health, Peking University, Beijing, China. No.38, Xueyuan Rd, Haidian District, Beijing 100191, China.
Email: liumin@bjmu.edu.cn

Funding information

National Key Research and Development Program of China, Grant/Award Number: 2021ZD0114104

Abstract

Mpox outbreak globally during 2022–2023, with more than 90% of cases occurring in men who have sex with men (MSM). However, the spatiotemporal distribution of mpox is not well established yet. This study aimed to explore the spatiotemporal clustering of mpox cases in MSM worldwide. We obtained the numbers of mpox cases from Our World in Data, the number of MSM from the Joint United Nations Programme on HIV/AIDS (UNAIDS), UNAIDS DATA 2021 and UNAIDS Global AIDS Update 2022 and literature. We evaluated the spatiotemporal cluster of mpox in MSM using retrospective space–time analyses method. The total number of mpox cases was 85 795 during May 1, 2022 to March 31, 2023. The most likely cluster was Spain (likelihood ratio = 4764.97; $p < 0.001$), with a cluster period from July 26 to August 14, 2022. There were 11 secondary clusters, which included 46 countries located in western Europe, eastern and northern South America, northern Europe, Canada, Central Africa, southern and central Europe, Latin America, Turkey, Dominican Republic, New Zealand, and Australia. The findings may inform current and future control strategies of mpox and might provide references for the identification of the spatiotemporal distribution of new and emerging infectious diseases in specific populations.

KEYWORDS

epidemiology, mpox, MSM, spatiotemporal analysis

1 | INTRODUCTION

Mpox (formerly known as monkeypox) is a zoonotic infectious disease. The virus includes two branches, the Central African branch and the West African branch. The current epidemic in the non-African region was caused by the West African branch, which has an incubation period of 5–21 days (typically 6–13 days),¹ and a mortality rate of 1.6%.^{2,3} The mpox virus can be transmitted through airborne droplets, direct or indirect contact with lesions or contaminated

objects,⁴ and vertical transmission.⁵ Whether it can be sexually transmitted remains to be validated.⁴

The number of confirmed and suspected cases of mpox has increased over the past 50 years (1970–2021) from 47 cases in 1970–1979 to 19 068 cases in 2010–2019.⁶ Cases were previously concentrated in Central and West Africa.⁷ However, recently, some travel-associated cases have been reported in Europe, the Americas, Asia, and Oceania.⁸ The number of cases in Europe has increased rapidly since May 2022 after initial mpox cases were reported in the

United Kingdom.^{1,9} On July 2022, the WHO declared mpox a public health emergency of international concern.¹⁰ As of April 4, 2023, there were 86 838 cases and 112 deaths¹¹ in 110 countries and regions over a period of 11 months, with most cases occurring in Europe, North America, and South America. Notably, more than 90% of cases were reported in men who have sex with men (MSM).¹

In the field of infectious diseases, spatiotemporal distribution analysis refers to the statistical analysis used to estimate distribution characteristics and the change pattern of an infectious disease over time and space. Due to more frequent risk behaviors such as sexual contact and multiple sexual partners, MSM are under the increased risk of mpox infection. Therefore, the spatiotemporal distribution analysis of mpox in MSM is vital in assessing and monitoring mpox's occurrence, intensity and direction of transmissibility,¹² but few studies estimate the spatiotemporal distribution in MSM from a global perspective because mpox previously occurred only in parts of Africa and did not draw considerable attention from other countries and regions.

In addition, despite the overall declining trend of mpox prevalence from 2022 to 2023, there still exists small-scale outbreaks globally with a slightly increased number of weekly reported new cases in parts of the world.¹³ Up to now, some studies were conducted only based on cases or case series of mpox.^{4,14,15} For example, one study analyzed geographical clusters of cumulative mpox cases and its virus lineages by countries and regions between September 2018 and August 2022.¹³ Another study analyzed the spatial distribution of mpox cases and its changes by months at global level.¹⁶ Mandja et al.¹² found spatiotemporal clusters in suspected or confirmed cases of 292 in the Democratic Republic of the Congo between 2000 and 2015. However, the above three studies did not focus on the spatiotemporal distribution of mpox in MSM. Using national surveillance data to analyze mpox cluster is one of the alternative methods, we should note that not all countries have available and high-quality surveillance data, which limits the description of mpox spatiotemporal cluster among different countries worldwide.

At present, several methods, including ClusterSeer, GeoSurveillance, kernel density, SanTScan and Flex Scan have been applied to detect and validate spatiotemporal aggregation of infectious diseases. Among them, SanTScan has a higher sensitivity for aggregate cluster analysis than others.^{17,18} Thus, in this study, we adopted SanTScan to determine whether there is a spatiotemporal cluster of mpox in MSM among 83 countries. Our study might provide insights into the spatiotemporal clusters of mpox among different countries from a global perspective and provide a reference for the development of prevention and control strategy in high-risk populations.

2 | METHODS

2.1 | Data source

We searched Our World in Data¹⁹ to obtain the data on daily number of new cases and cumulative cases of mpox in 83 countries between

May 1, 2022 and March 31, 2023. Our World in Data is an open global database, which covers most countries and focuses on poverty, disease, hunger, climate change, war, existential risk and inequality.¹⁹ Data on mpox outbreak are collated from WHO, updated every hour, and kept up to the previous day. All data are available on GitHub.¹⁹ We used May 1, 2022 as the start time because the data reported in Our World in Data began on May 1, 2022. We used March 31, 2023 as the end time because the data were updated to March 31, 2023 when we performed this study. Data on the longitude and latitude of each country were obtained from Model Whale (Supporting Information: Table S1).²⁰ MSM data were collected from the Joint United Nations Programme on HIV/AIDS (UNAIDS),²¹ UNAIDS DATA 2021 and UNAIDS Global AIDS Update 2022, and literature (Supporting Information: Table S2). The number of MSM in 83 countries was available. In our study, MSM include the men who have sex with men and women, and the men who are exclusively gay/same-sex attracted and only have sex with other gay men.²²

2.2 | Statistical analysis

The daily new cases of mpox in North America, Europe, South America, Africa, Asia, and Oceania were calculated and presented as a bar chart. The cumulative confirmed cases and MSM population were described at country levels.

The retrospective space-time analysis method based on discrete Poisson model was used to identify the potential spatiotemporal cluster of mpox in MSM.^{23,24} In this analysis, to evaluate the change in a number of cases inside and outside the window, a scanning window was constructed in the form of a cylinder, whose height represents time and base area represents region. The position of scanning center was selected randomly. The height and base area were constantly changing until all the space units were scanned.²⁵ We calculated the percentage of mpox cases in MSM population in each country and found that 3.3% was the highest prevalence rate among 83 countries. Thus, we set 3.3% as the upper limit of geographic size in the scanning window to ensure that at least one spatiotemporal cluster could be detected among 83 countries. To detect the clusters of mpox at different incubation periods, we set the time length of scanning window for 7, 14, and 21 days, respectively, according to the shortest, median, and longest incubation periods of mpox. Log-likelihood ratio (LLR) was calculated using the observed and theoretical number of cases inside and outside the window. The largest LLR value indicated the most likely cluster area. Additionally, other windows with statistical significance of LLR indicated secondary clusters. The Monte Carlo approach was used to test the statistical significance of the LLR. Relative risk (RR) was calculated to evaluate the strength of aggregation.²⁶ The *p* value of less than 0.05 indicated statistical significance. The details regarding the hypothesis test and the calculation of LLR and RR were listed below.

Null hypothesis (H0): The spatial and temporal distribution of mpox in MSM is random;

Alternative hypothesis (H1): The spatial and temporal distribution of mpox in MSM is not random.

$$E(c) = \frac{C}{P} \times p,$$

where c is the observed number of cases and p is the number of MSM population in the region within the window, while C and P are the total number of mpox cases and MSM population respectively. $E[c]$ is the theoretical number of cases within the window under the null-hypothesis. LLR and RR were calculated as below:

$$\text{LLR} = \left(\frac{c}{E[c]} \right)^c \left(\frac{C-c}{C-E[c]} \right)^{C-c}$$

$$\text{RR} = \frac{c/E[c]}{(C-c)/(C-E[c])}$$

Microsoft Excel was used to collate the data. Line graphs were plotted to show the changes in the daily numbers of cases in six continents. SaTScan (version 10.1; developed by Kulldorff; <https://www.satscan.org/>) was used to analyze the spatiotemporal clusters of mpox cases in MSM. ArcGIS (version 9.4) was used to visualize the results.

2.3 | Ethical approval

Ethical approval is not required for this study, given that the study does not involve direct data collection from people.

3 | RESULTS

3.1 | Prevalence of mpox

The total number of mpox cases in the 83 countries was 85 795 between May 1, 2022 and March 31, 2023, and the total estimated number of MSM population was 28 625 546. The United States had the highest number of cases (30 079) between June 3, 2022 and March 31, 2023. The estimated number of MSM in the United States was 4 604 040 during the same period. The second and third total numbers of cases in Brazil and Spain were 10 890 and 7546, and the estimated numbers of MSM were 2 000 000 and 890 200, respectively. The top five countries with the highest incidence rate of mpox were Luxembourg (3259.01 per 100 000 MSM population), Costa Rica (2148.51 per 100 000 MSM population), Peru (1455.77 per 100 000 MSM population), France (1250.91 per 100 000 MSM population), and Chile (1168.29 per 100 000 MSM population) (Table 1).

Figure 1 presented the number of daily cases and changing trend of mpox in six continents. Five continents including North America, Europe, South America, Africa, and Oceania had a similar trend that showed an increase followed by a decrease in the number of daily

cases, with one peak observed during the period. However, despite a small number of daily cases in Asia, there was a remarkably increased trend of daily cases after mid-March 2023.

3.2 | Spatiotemporal cluster of mpox in MSM

To detect the clusters of mpox at different incubation periods, we set the time length of scanning window for 7, 14, and 21 days, respectively, according to the shortest, median, and longest incubation periods of mpox.

When the temporal window was set at 7 days, we identified one most likely spatiotemporal cluster and 15 secondary clusters, which covered a total of 47 countries. The most likely spatiotemporal cluster was Spain (LLR = 2834.13, $p < 0.001$, cluster period: July 5–10, 2022). The first secondary cluster was France (LLR = 2282.69, $p < 0.001$, cluster time: July 21, 2022), the second secondary cluster was Ecuador, Colombia and Peru (LLR = 1697.58, $p < 0.001$, cluster time: October 12, 2022), and the third secondary cluster was Ireland, the UK and Netherlands (LLR = 1234.23, $p < 0.001$, cluster period: July 18–19, 2022). The other secondary clusters are shown in Figure 2A and Supporting Information: Table S3.

When the temporal window was set at 14 days and 21 days, we found a similar result. There was one most likely spatiotemporal cluster and 11 secondary clusters, which also covered 47 countries. The most likely cluster was Spain (14 days: LLR = 4008.34, $p < 0.001$, cluster period: July 5–18, 2022; 21 days: LLR = 4764.97, $p < 0.001$ cluster period: July 26 to August 14, 2022). The first secondary cluster was France, Switzerland, Luxembourg, Belgium and Netherlands (14 days: LLR = 3263.26, $p < 0.001$, cluster period: July 21 to August 2, 2022; 21 days: LLR = 4244.08, $p < 0.001$, cluster period: July 19 to August 8, 2022); the second secondary cluster was in Peru, Ecuador, Bolivia and Colombia (14 days: LLR = 2671.25, $p < 0.001$, cluster period: September 20–28, 2022; 21 days: LLR = 3641.44, $p < 0.001$, cluster period: September 20 to October 5, 2022), and the third secondary cluster was Norway, Denmark, Sweden and the UK (14 days: LLR = 1324.69, $p < 0.001$, cluster period: July 6–19, 2022; 21 days: LLR = 1636.05, $p < 0.001$, cluster period: July 18 to August 2, 2022). The other secondary clusters were shown in Figure 2B,C, Supporting Information: Tables S4 and Table S5.

Figure 3 presented the RR values for spatiotemporal clusters across different temporal windows. When the temporal window was set at 7 days, we observed one largest RR value in the second secondary cluster including France (RR = 184.40, $p < 0.001$). The second largest RR value was in the twelfth secondary cluster including Turkey (RR = 178.47, $p < 0.001$), and the third largest RR value was in the third secondary cluster including Ecuador, Colombia, Peru (RR = 79.73, $p < 0.001$) (Supporting Information: Table S3 and Figure 3A).

Notably, when the temporal window was set at 14 days and 21 days, we found a similar result. The values of top two RR were same, and the largest RR value was in the eighth secondary cluster, including Turkey (RR = 178.47, $p < 0.001$), and the second largest RR value was in 10th secondary cluster including New Zealand (RR =

TABLE 1 Cumulative cases of mpox and number of the MSM in 83 countries from May 1, 2022 to March 31, 2023.

| Country | Period | Cumulative mpox cases | MSM population | Incidence of mpox per 100 000 |
|----------------------------------|----------------------|-----------------------|----------------|-------------------------------|
| Argentina | 2022/6/3–2023/3/31 | 1124 | 205 600 | 546.69 |
| Australia | 2022/5/20–2022/12/9 | 144 | 263 500 | 54.65 |
| Bahamas | 2022/6/27–2023/3/31 | 2 | 2800 | 71.43 |
| Barbados | 2022/7/19–2023/3/31 | 1 | 2600 | 38.46 |
| Belgium | 2022/5/19–2023/3/28 | 793 | 144 753 | 547.83 |
| Benin | 2022/6/24–2023/3/24 | 3 | 5800 | 51.72 |
| Bolivia | 2022/8/3–2023/3/31 | 265 | 35 500 | 746.48 |
| Bosnia and Herzegovina | 2022/7/14–2023/3/28 | 9 | 6900 | 130.43 |
| Brazil | 2022/6/10–2023/3/31 | 10 890 | 2 000 000 | 544.50 |
| Bulgaria | 2022/6/23–2023/3/28 | 6 | 57 800 | 10.38 |
| Cameroon | 2022/5/1–2023/3/24 | 14 | 7000 | 200.00 |
| Canada | 2022/6/3–2023/3/31 | 1478 | 349 800 | 422.53 |
| Central African Republic | 2022/5/1–2023/3/24 | 23 | 3000 | 766.67 |
| Chile | 2022/6/18–2023/3/31 | 1437 | 123 000 | 1168.29 |
| China | 2022/6/24–2023/3/28 | 24 | 8 288 536 | 0.29 |
| Colombia | 2022/6/25–2023/3/31 | 4089 | 357 000 | 1145.38 |
| Republic of Congo | 2022/5/1–2023/3/24 | 3 | 1300 | 230.77 |
| Costa Rica | 2022/7/21–2023/3/31 | 217 | 10 100 | 2148.51 |
| Croatia | 2022/6/24–2023/3/28 | 33 | 29 500 | 111.86 |
| Cuba | 2022/8/23–2023/3/31 | 8 | 279 200 | 2.87 |
| Czech Republic | 2022/5/24–2023/3/28 | 71 | 109 644 | 64.76 |
| Democratic Republic of the Congo | 2022/5/20–2023/3/24 | 439 | 194 900 | 225.24 |
| Denmark | 2022/5/23–2023/3/28 | 196 | 50 000 | 392.00 |
| Dominican Republic | 2022/7/7–2023/3/31 | 52 | 142 000 | 36.62 |
| Ecuador | 2022/7/6–2023/3/31 | 530 | 89 400 | 592.84 |
| Egypt | 2022/9/27–2022/12/12 | 3 | 64 300 | 4.67 |
| El Salvador | 2022/9/1–2023/3/31 | 98 | 54 100 | 181.15 |
| Estonia | 2022/6/28–2023/3/28 | 11 | 9000 | 122.22 |
| France | 2022/5/19–2023/3/28 | 4128 | 330 000 | 1250.91 |
| Georgia | 2022/6/15–2023/3/28 | 2 | 19 000 | 10.53 |
| Germany | 2022/5/20–2023/3/28 | 3692 | 750 000 | 492.27 |
| Ghana | 2022/5/1–2023/3/24 | 122 | 54 800 | 222.63 |
| Greece | 2022/6/9–2023/3/28 | 87 | 94 000 | 92.55 |
| Guatemala | 2022/8/4–2023/3/31 | 404 | 116 500 | 346.78 |
| Guyana | 2022/8/24–2023/3/31 | 2 | 3300 | 60.61 |
| Honduras | 2022/8/14–2023/3/31 | 40 | 40 900 | 97.80 |
| Hungary | 2022/5/31–2023/3/28 | 80 | 53 404 | 149.80 |
| India | 2022/7/14–2023/3/25 | 22 | 238 200 | 9.24 |

TABLE 1 (Continued)

| Country | Period | Cumulative mpx cases | MSM population | Incidence of mpx per 100 000 |
|--------------|---------------------|----------------------|----------------|------------------------------|
| Indonesia | 2022/8/22–2023/3/25 | 1 | 754 300 | 0.13 |
| Iran | 2022/8/18 | 1 | 359 000 | 0.28 |
| Ireland | 2022/5/16–2023/3/28 | 228 | 86 500 | 263.58 |
| Italy | 2022/5/19–2023/3/28 | 957 | 359 315 | 266.34 |
| Jamaica | 2022/7/8–2023/3/31 | 21 | 42 400 | 49.53 |
| Japan | 2022/5/1–2023/3/30 | 82 | 700 000 | 11.71 |
| Latvia | 2022/6/3–2023/3/28 | 6 | 12 880 | 46.58 |
| Lebanon | 2022/6/20–2023/3/9 | 27 | 17 000 | 158.82 |
| Liberia | 2022/7/29–2023/3/24 | 10 | 74 600 | 13.40 |
| Lithuania | 2022/8/4–2023/3/28 | 5 | 17 760 | 28.15 |
| Luxembourg | 2022/6/16–2023/3/28 | 57 | 1749 | 3259.01 |
| Mexico | 2022/6/3–2023/3/31 | 3937 | 1 226 000 | 321.13 |
| Moldova | 2022/8/9–2023/3/28 | 2 | 14 600 | 13.70 |
| Morocco | 2022/6/2–2022/8/29 | 3 | 42 000 | 7.14 |
| Mozambique | 2022/10/7–2023/3/24 | 1 | 15 800 | 6.33 |
| Netherlands | 2022/5/22–2023/3/28 | 1262 | 230 000 | 548.70 |
| New Zealand | 2022/7/11–2023/1/22 | 41 | 37 500 | 109.33 |
| Nigeria | 2022/5/1–2023/3/24 | 814 | 240 000 | 339.17 |
| Norway | 2022/5/31–2023/3/28 | 95 | 56 459 | 168.26 |
| Panama | 2022/7/6–2023/3/31 | 189 | 30 000 | 630.00 |
| Paraguay | 2022/8/26–2023/3/31 | 119 | 32 200 | 369.57 |
| Peru | 2022/6/28–2023/3/31 | 3785 | 260 000 | 1455.77 |
| Philippines | 2022/7/29–2022/8/22 | 4 | 687 100 | 0.58 |
| Poland | 2022/6/13–2023/3/28 | 215 | 67 482 | 318.60 |
| Portugal | 2022/5/17–2023/3/28 | 951 | 103 153 | 921.93 |
| Romania | 2022/6/14–2023/3/28 | 47 | 10 500 | 447.62 |
| Russia | 2022/7/12–2023/3/28 | 2 | 243 384 | 0.82 |
| Serbia | 2022/6/17–2023/3/28 | 40 | 40 000 | 100.00 |
| Singapore | 2022/6/21–2023/3/30 | 22 | 220 000 | 10.00 |
| Slovakia | 2022/7/7–2023/3/28 | 14 | 18 614 | 75.21 |
| Slovenia | 2022/5/24–2023/3/28 | 47 | 39 427 | 119.21 |
| South Africa | 2022/7/1–2023/3/24 | 5 | 310 000 | 1.61 |
| Spain | 2022/5/18–2023/3/28 | 7546 | 890 200 | 847.67 |
| Sri Lanka | 2022/11/4–2023/3/25 | 2 | 74 000 | 2.70 |
| Sudan | 2022/8/1–2022/10/19 | 18 | 132 000 | 13.64 |
| Sweden | 2022/5/18–2023/3/28 | 260 | 100 000 | 260.00 |
| Switzerland | 2022/5/21–2023/3/28 | 552 | 80 000 | 690.00 |
| Thailand | 2022/7/21–2023/3/25 | 18 | 527 900 | 3.41 |

(Continues)

TABLE 1 (Continued)

| Country | Period | Cumulative mpx cases | MSM population | Incidence of mpx per 100 000 |
|--------------------|----------------------|----------------------|----------------|------------------------------|
| Turkey | 2022/6/30–2023/3/28 | 12 | 6890 | 174.17 |
| Ukraine | 2022/9/15–2023/3/28 | 5 | 180 000 | 2.78 |
| The United Kingdom | 2022/5/7–2023/3/28 | 3738 | 598 256 | 624.82 |
| The United States | 2022/6/3–2023/3/31 | 30 079 | 4 604 040 | 653.32 |
| Uruguay | 2022/7/31–2023/3/31 | 19 | 28 600 | 66.43 |
| Venezuela | 2022/6/15–2023/3/31 | 12 | 210 800 | 5.69 |
| Vietnam | 2022/10/3–2022/10/19 | 2 | 256 000 | 0.78 |
| Total | 2022/5/1–2023/3/31 | 85 795 | 28 625 546 | 299.71 |

Abbreviations: mpx, monkeypox; MSM, men who have sex with men.

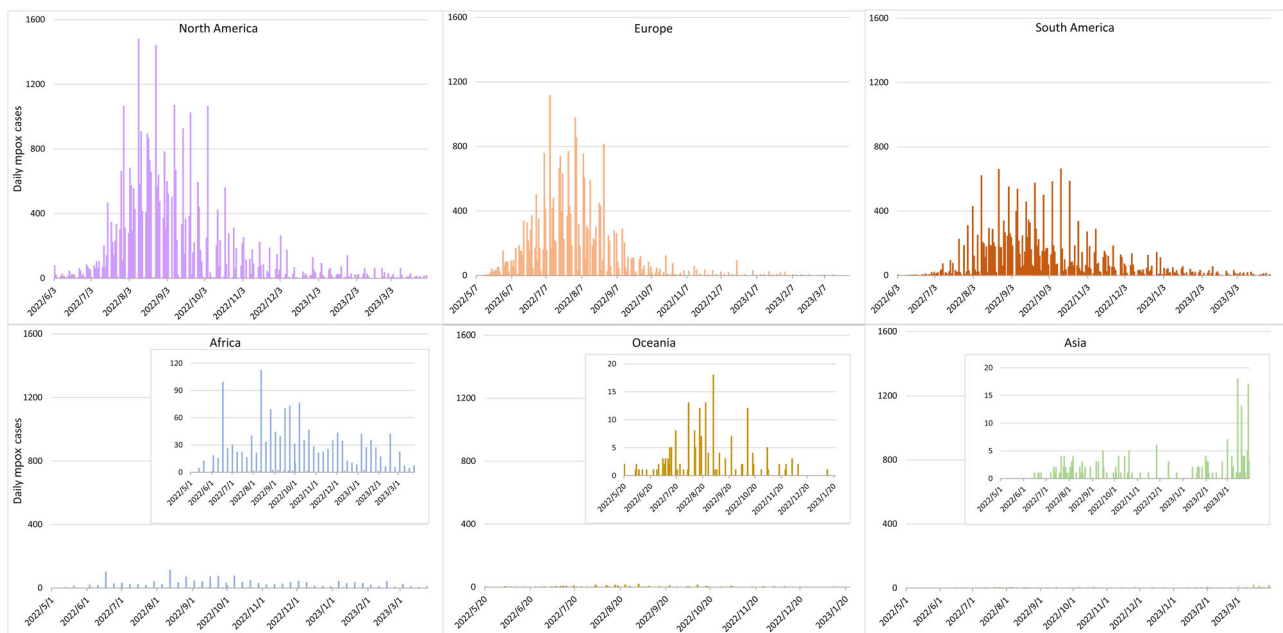


FIGURE 1 The number of daily cases of mpx in six continents between May 1, 2022 and March 31, 2023. mpx, monkeypox.

35.77, $p < 0.001$). The third largest RR value was in fourth secondary cluster at 14 days and fifth secondary cluster at 21 days including Liberia, Ghana, Benin, Nigeria, Cameroon, Morocco, Republic of Congo, Central African Republic, Democratic Republic of the Congo, and Portugal (RR = 26.51, $p < 0.001$) (Supporting Information: Tables S4, S5, and Figure 3B,C). RR corresponding to each country in the spatiotemporal cluster of mpx from May 1, 2022 to March 31, 2023 were shown in Supporting Information: Table S6.

4 | DISCUSSIONS

We analyzed the prevalence and the clusters of mpx cases in MSM using a retrospective spatiotemporal model and a geographic information system. A total of 85 795 mpx cases were reported in

83 countries between May 1, 2022 and March 31, 2023. We discovered 12 spatiotemporal clusters of mpx cases in MSM covering 47 countries. The most likely cluster was Spain, and the other 11 secondary clusters located in western Europe, eastern and northern South America, northern Europe, Canada, Central Africa, Southern and central Europe, Latin America, Turkey, Dominican Republic, New Zealand and Australia. To the best of our knowledge, this is the first study to explore the cluster of mpx cases in MSM. The findings of the present study may improve the understanding of the global spatiotemporal epidemiology of mpx in MSM.

We observed that mpx cases were clustered in MSM. Most mpx outbreaks in western and southern European countries, such as Spain, Portugal, Italy and the UK, were initially reported in MSM population,^{1,9,27} and the outbreaks in countries located in North America, Latin America, and Oceania, such as Canada, and Australia,

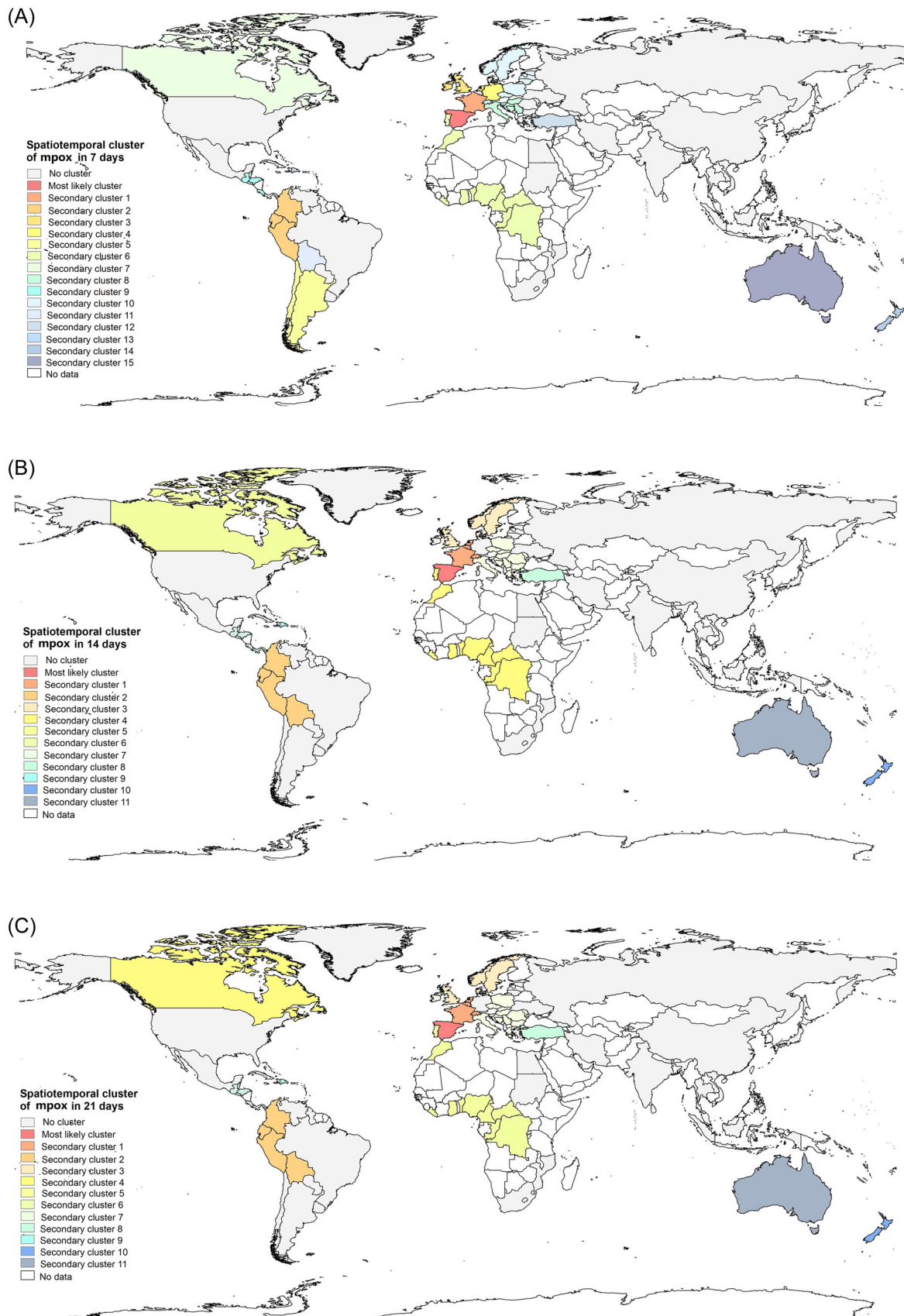


FIGURE 2 Spatiotemporal clusters of mpx in MSM population in 83 countries from May 1, 2022 to March 31, 2023. (A) Spatiotemporal clusters of mpx in 7 days; (B) spatiotemporal clusters of mpx in 14 days; and (C) spatiotemporal clusters of mpx in 21 days. mpx, monkeypox.

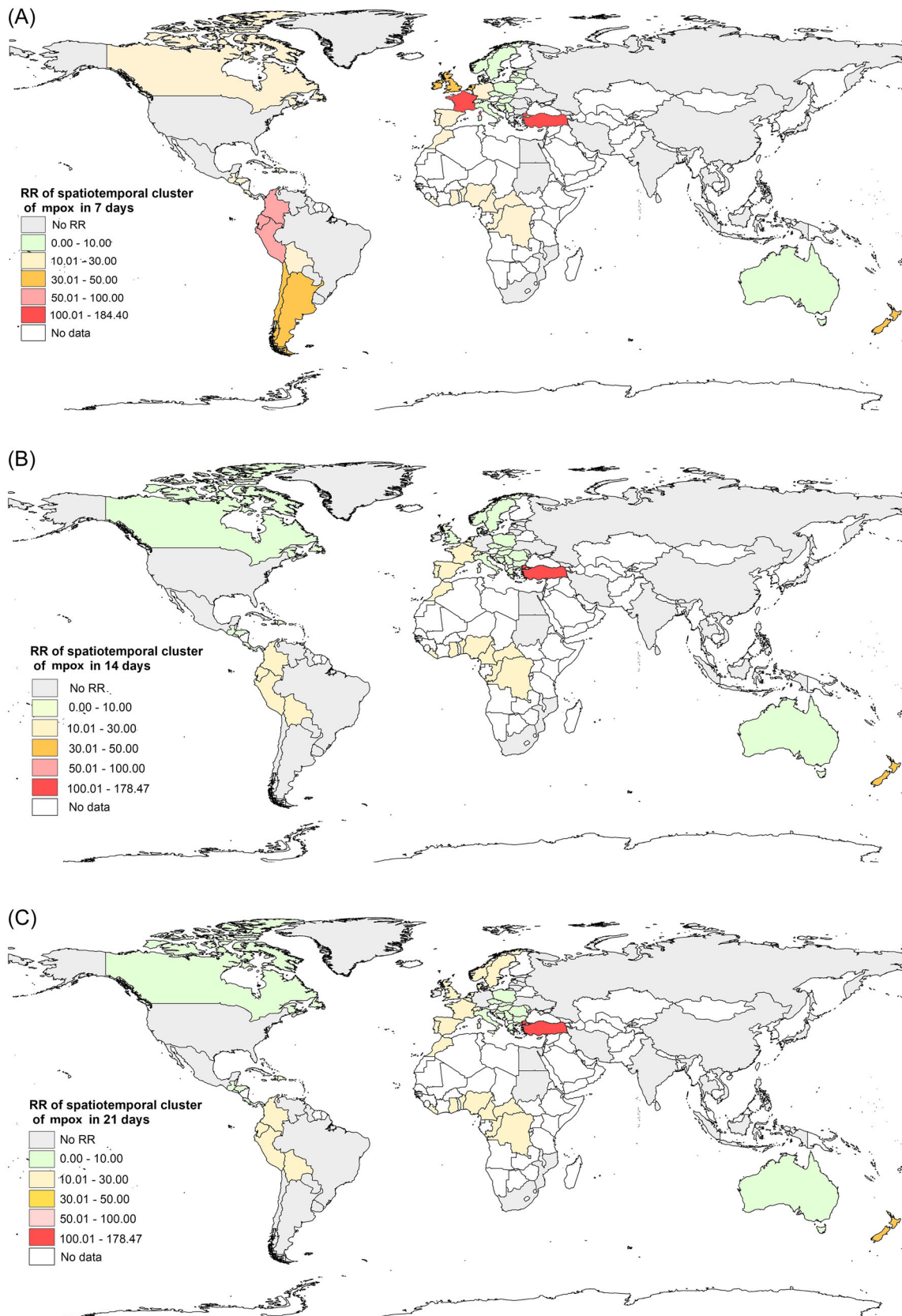


FIGURE 3 Relative risk of spatiotemporal clusters of mpox in MSM population from May 1, 2022 to March 31, 2023 in 83 countries. (A) Relative risk of spatiotemporal clusters of mpox in 7 days; (B) relative risk of spatiotemporal clusters of mpox in 14 days; (C) relative risk of spatiotemporal clusters of mpox in 21 days. mpox, monkeypox; RR, relative risk.

also disproportionately affected MSM.^{8,28} The early transmission of the mpox virus in cases demonstrates a high correlation in sexual networks within the male population,²⁹ and more than 90% of reported cases are MSM.² Previous studies showed that the R_0 of mpox is 2.43 in MSM.³⁰ Owing to the transmission of skin-to-skin contact, the clustered spread of mpox in MSM can be caused by risk behaviors, such as condomless penile–anal intercourse, multiple sexual partners, and sexual encounters in bars, clubs, and other parties.^{31,32} The infected individuals will increase the spread of mpox through sexual contact behavior with his sexual partners. If a gay man has many sexual partners, a clustered outbreak of mpox may happen among MSM. Countries with cultural tolerance of homosexuality are more likely to have a significant increase in mpox cases in gay communities in a short time.³³ In contrast, in homosexuality-prohibiting countries,³⁴ a gay man has to marry a woman and have sexual contact simultaneously with his wife and male sexual partner. Once the gay man infected mpox, the risk of clustering in the family and sexual networks increase. However, Africa countries have a lower level of medical care, and their healthcare workers lack knowledge of managing the sexual health needs of MSM population,³⁵ detection technology, diagnostic skills and treatment capacity comparing to developed countries. Even in the UK, a developed country, some cases were misdiagnosed as infections caused by the herpes simplex virus or varicella-zoster virus due to the same clinical manifestations of mpox and sexually transmitted diseases.³⁶ Thus, the undiagnosed cases, delayed diagnosis, and insufficient treatment in MSM can lead to the outbreak and even the epidemic of mpox.

Although mpox is a self-limiting disease, with a low rate of associated mortality, patients with this disease are vulnerable to social stigma.³⁷ Smallpox vaccines can be used to control mpox outbreaks, whereas the first-generation smallpox vaccines in national reserves are not recommend as they do not meet the current safety and manufacturing standards. The second-generation vaccine, ACAM2000, had limited safety data from large population-based program and was associated with rare but serious adverse event, such as myopericarditis.³⁸ JYNNEOS is a third-generation vaccine, which was authorized to manufacture by the US FDA on 24 September 2019.³⁹ Since August 9, 2022, under an Emergency Use Authorization by FDA, the standard regimen has been authorized for people aged <18 years, and an alternative regimen used for people age ≥ 18 years.⁴⁰ The JYNNEOS vaccine can reduce the risk of mpox infection and the incidence of severe illness, with fewer adverse events occurring.⁴¹ However, this vaccine is only available in parts of countries. In addition to vaccination, other specifically clinical treatment for mpox is still under development.² Moreover, some strategies target on mpox prevention can also reduce the risk of mpox infection, including restricted number of sexual partners, reduction in sexual acts frequency, avoidance of sexual intercourse with temporary partners from dating apps or sex venues, adoption of condoms and disinfection measures.^{31,42} A study showed that more than 55% of MSM and transgender women adopted these strategies.³¹

The most likely spatiotemporal cluster of mpox in MSM was in Spain, and two possible reasons could explain the phenomenon. First, a celebration, Gay Pride Maspalomas festival, held in Gran Canaria

during May 5–15, 2022 might be the origin of the outbreak of mpox. Around 30 000 overseas visitors, who were lesbian, gay, bisexual, and transgender (LGBT), attended this gathering and some of them had high number of sexual partners during their stay.⁴³ After 2 days (May 17, 2022), seven mpox cases were first observed in a sexually transmitted disease clinic in Madrid of Spain, and all cases have no identified epidemiological links to mpox cases in other countries.³² It was 30 days later (June 22, 2022) that 508 confirmed cases occurred in Madrid and almost of them are MSM. Some of cases attended the Gay Pride Maspalomas festival or attended a same sauna and were exposed to condomless sex with unknown partners.³² In terms of time, the mass outbreak of cases coincided with the longest incubation period of mpox. Epidemiological links between cases may accelerate the outbreak of mpox in MSM in Spain. Second, an inclusive social environment may facilitate the sustainable spread of mpox among MSM populations. The marriage of people between same sex was legalized in Spain since 2005, and more than 80% Spanish show an acceptable attitude toward homosexuality.⁴⁴ Notably, Madrid is a tolerant LGBT city, with a vibrant gay and other men who have sex with men community.³³ All above factors could be potential important reasons for the outbreak and prevalence of mpox in MSM in Spain. In addition, MSM population have access to public health services, such as sexual health test, by the way of privacy and free of charge in Spain.⁴⁵ This might be helpful for MSM to detect mpox virus in the early period.

Spatiotemporal aggregation of mpox in MSM populations obtained using space-time statistical model provided additional epidemiological information of the disease. Identifying the areas with high spatiotemporal aggregation of cases and understanding cluster situation may improve in public health control measures. Retrospective scan statistics has been recognized as one of the most comprehensive methods to evaluate the spatial and temporal distribution of infectious diseases, such as tuberculosis, Malaria, Covid-19, and so forth.^{17,26,46} Previous studies showed that this method have higher sensitivity to detect true infectious disease (Malaria) clusters compared with other methods.¹⁷

The cases of mpox are currently low in many countries and regions, however, those countries have high numbers of MSM, which may increase the likelihood of transmission. Therefore, in countries where mpox is endemic, proactive measures such as the screening of high-risk populations and timely detection of confirmed cases are required to control the viral transmission.

Our study has several limitations. First, data on the MSM populations of the included countries were obtained from multiple sources and corresponded to different years, thus the possibility of the overestimation or underestimation of the risk of clustering cannot be ignored. Second, for cluster analysis, we collected data from Our World in Data. The data did not include the number of confirmed mpox cases in MSM. Studies have indicated that approximately 98% of all confirmed cases of mpox were reported in MSM.^{4,9} Thus, the spatiotemporal cluster results obtained using the data regarding the total number of confirmed cases might have been overestimated. Third, the extrapolation of this study should be cautious because the countries included this study cannot represent all countries in global.

Fourth, scan statistics method is limited by the circular search window and may not capture irregularly or noncircular clusters. Besides, its sensitivity may be affected by the choice of scan window size and the significance level used for cluster detection.

In conclusion, 12 spatiotemporal clusters of mpox cases were found in MSM population and the most likely cluster was in Spain, the secondary clusters were in western Europe, eastern and northern South America, northern Europe, Canada, Central Africa, Southern and central Europe, Latin America, Turkey, Dominican Republic, New Zealand and Australia. The study might inform current and future control strategies of mpox, and enhance the identification of the spatiotemporal distribution of new and emerging infectious diseases in specific population.

AUTHOR CONTRIBUTIONS

Min Liu, Jue Liu, and Weijing Shang conceived and designed the study. Guiying Cao, Yu Wu, Liangyu Kang, Yaping Wang, Peng Gao, and Weijing Shang collected, analyzed and interpreted the data. Weijing Shang, Guiying Cao, Yu Wu, Yaping Wang, and Peng Gao wrote the manuscript. Guiying Cao, Liangyu Kang, Jue Liu, and Min Liu critically revised the manuscript. Min Liu supervised the study, and have full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. All authors reviewed the study findings and read and approved the final version before submission.

ACKNOWLEDGMENTS

We appreciate the works by the Our World in Data, Model Whale, and the Joint United Nations Programme on HIV/AIDS. This work was supported by grants 2021ZD0114104 from the National Key Research and Development Program of China.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in Our World in Data at <https://ourworldindata.org/monkeypox>. These data were derived from the following resources available in the public domain: <https://ourworldindata.org/monkeypox>, <https://kpatlas.unaids.org/dashboard>. The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Jue Liu  <http://orcid.org/0000-0002-1938-9365>

Min Liu  <http://orcid.org/0000-0002-5059-3743>

REFERENCES

- Singhal T, Kabra SK, Lodha R. Monkeypox: a review. *Indian J Pediatr.* 2022;89(10):955-960. doi:10.1007/s12098-022-04348-0
- World Health Organization. Monkeypox. 2022. Accessed November 28, 2022. <https://www.who.int/news-room/fact-sheets/detail/monkeypox>
- Branda F, Pierini M, Mazzoli S. Monkeypox: early estimation of basic reproduction number $R(0)$ in Europe. *J Med Virol.* 2022;95(1):e28270. doi:10.1002/jmv.28270
- Thornhill JP, Barkati S, Walmsley S, et al. Monkeypox virus infection in humans across 16 countries - April-June 2022. *N Engl J Med.* 2022;387(8):679-691. doi:10.1056/NEJMoa2207323
- Lapa D, Carletti F, Mazzotta V, et al. Monkeypox virus isolation from a semen sample collected in the early phase of infection in a patient with prolonged seminal viral shedding. *Lancet Infect Dis.* 2022;22(9):1267-1269. doi:10.1016/s1473-3099(22)00513-8
- Bunge EM, Hoet B, Chen L, et al. The changing epidemiology of human monkeypox—a potential threat? A systematic review. *PLoS Neglected Trop Dis.* 2022;16(2):e0010141. doi:10.1371/journal.pntd.0010141
- Yang Z. Monkeypox: a potential global threat. *J Med Virol.* 2022;94(9):4034-4036. doi:10.1002/jmv.27884
- Adalja A, Inglesby T. A novel international monkeypox outbreak. *Ann Intern Med.* 2022;175(8):1175-1176. doi:10.7326/m22-1581
- Gessain A, Nakoune E, Yazdanpanah Y. Monkeypox. *N Engl J Med.* 2022;387(19):1783-1793. doi:10.1056/NEJMra2208860
- World Health Organization. WHO Director-General's statement at the press conference following IHR Emergency Committee regarding the multi-country outbreak of monkeypox - 23 July 2022. 2022. Accessed November 28, 2022. <https://www.who.int/news-room/speeches/item/who-director-general-s-statement-on-the-press-conference-following-ihr-emergency-committee-regarding-the-multi-country-outbreak-of-monkeypox-23-july-2022>
- World Health Organization. 2022-23 Mpox (Monkeypox) Outbreak: Global Trends. Accessed April 7, 2023. https://worldhealthorg.shinyapps.io/mpx_global/
- Mandja BAM, Brembilla A, Handschumacher P, et al. Temporal and spatial dynamics of monkeypox in democratic Republic of Congo, 2000-2015. *EcoHealth.* 2019;16(3):476-487. doi:10.1007/s10393-019-01435-1
- Chakraborty C, Bhattacharya M, Sharma AR, Dhama K. Evolution, epidemiology, geographical distribution, and mutational landscape of newly emerging monkeypox virus. *GeroScience.* 2022;44(6):2895-2911. doi:10.1007/s11357-022-00659-4
- Philpott D, Hughes CM, Alroy KA, et al. Epidemiologic and clinical characteristics of monkeypox cases - United States, May 17-July 22, 2022. *MMWR Morb Mortal Wkly Rep.* 2022;71(32):1018-1022. doi:10.15585/mmwr.mm7132e3
- Vivancos R, Anderson C, Blomquist P, et al. Community transmission of monkeypox in the United Kingdom, April to May 2022. *Euro Surveill.* 2022;27(22):2200422. doi:10.2807/1560-7917.Es.2022.27.22.2200422
- Patwary MM, Hossain J, Billah SM, Kabir MP, Rodriguez-Morales AJ. Mapping spatio-temporal distribution of monkeypox disease incidence: a global hotspot analysis. *New Microb New Infect.* 2023;53:101150. doi:10.1016/j.nmni.2023.101150
- Gwitira I, Mukonoweshuro M, Mapako G, Shekede MD, Chirenda J, Mberikunasho J. Spatial and spatio-temporal analysis of malaria cases in Zimbabwe. *Infect Dis Poverty.* 2020;9(1):146. doi:10.1186/s40249-020-00764-6
- Barro AS, Kracalik IT, Malania L, et al. Identifying hotspots of human anthrax transmission using three local clustering techniques. *Appl Geogr.* 2015;60:29-36. doi:10.1016/j.apgeog.2015.02.014
- Our World in Data. Mpox (monkeypox). 2022. Accessed November 16, 2022. <https://ourworldindata.org/monkeypox>
- Model Whale. Longitude and latitude of all countries in the world. Accessed November 17, 2022. <https://www.heywhale.com/mw/dataset/6051ae725316950016ee73f1/file>
- The Joint United Nations Programme on HIV/AIDS. UNAIDS Dashboard - Population size estimate (all ages) in men who have

- sex with men. Accessed November 16, 2022. <https://kpatlas.unaids.org/dashboard>
22. U.S. Center for Disease Control and Prevention. Men who have sex with men (MSM). Accessed September 3, 2023. <https://www.cdc.gov/std/treatment-guidelines/msm.htm>
 23. Greene SK, Peterson ER, Balan D, et al. Detecting COVID-19 clusters at high spatiotemporal resolution, New York City, New York, USA, June–July 2020. *Emerging Infect Dis.* 2021;27(5):1500–1504. doi:10.3201/eid2705.203583
 24. Rao H, Li DM, Zhao X, Yu J. Spatiotemporal clustering and meteorological factors affected scarlet fever incidence in mainland China from 2004 to 2017. *Sci Total Environ.* 2021;777:146145. doi:10.1016/j.scitotenv.2021.146145
 25. Zhu X, Zhu Z, Gu L, et al. Spatio-temporal variation on syphilis from 2005 to 2018 in Zhejiang Province, China. *Front Public Health.* 2022;10:873754. doi:10.3389/fpubh.2022.873754
 26. Zhao Y, Liu Q. Analysis of distribution characteristics of COVID-19 in America based on space-time scan statistic. *Front Public Health.* 2022;10:897784. doi:10.3389/fpubh.2022.897784
 27. Ward T, Christie R, Paton RS, Cumming F, Overton CE. Transmission dynamics of monkeypox in the United Kingdom: contact tracing study. *BMJ.* 2022;379:e073153. doi:10.1136/bmj-2022-073153
 28. Cabanillas B, Valdelvira R, Akdis CA. Monkeypox outbreak in Europe, UK, North America, and Australia: a changing trend of a zoonotic disease. *Allergy.* 2022;77(8):2284–2286. doi:10.1111/all.15393
 29. Bryer J, Freeman EE, Rosenbach M. Monkeypox emerges on a global scale: a historical review and dermatologic primer. *J Am Acad Dermatol.* 2022;87(5):1069–1074. doi:10.1016/j.jaad.2022.07.007
 30. Guzzetta G, Mammone A, Ferraro F, et al. Early estimates of monkeypox incubation period, generation time, and reproduction number, Italy, May–June 2022. *Emerg Infect Dis.* 2022;28(10):2078–2081. doi:10.3201/eid2810.221126
 31. Hubach RD, Owens C. Findings on the monkeypox exposure mitigation strategies employed by men who have sex with men and transgender women in the United States. *Arch Sex Behav.* 2022;51(8):3653–3658. doi:10.1007/s10508-022-02423-3
 32. Iñigo Martínez J, Gil Montalbán E, Jiménez Bueno S, et al. Monkeypox outbreak predominantly affecting men who have sex with men, Madrid, Spain, 26 April to 16 June 2022. *Euro Surveill.* 2022;27(27):2200471. doi:10.2807/1560-7917.ES.2022.27.27.2200471
 33. Santoro P, Rodríguez R, Morales P, Morano A, Morán M. One “chemsex” or many? Types of chemsex sessions among gay and other men who have sex with men in Madrid, Spain: findings from a qualitative study. *Int J Drug Policy.* 2020;82:102790. doi:10.1016/j.drugpo.2020.102790
 34. Shangani S, Genberg B, Harrison A, et al. Cultural adaptation and validation of a measure of prejudice against men who have sex with men among healthcare providers in Western Kenya. *Global Public Health.* 2022;17(1):150–164. doi:10.1080/17441692.2020.1860248
 35. Muwanguzi PA, Nabunya R, Karis VMS, Nabisere A, Nangendo J, Mujugira A. Nurses' reflections on caring for sexual and gender minorities pre-post stigma reduction training in Uganda. *BMC Nurs.* 2023;22(1):50. doi:10.1186/s12912-023-01208-w
 36. Heskin J, Belfield A, Milne C, et al. Transmission of monkeypox virus through sexual contact - a novel route of infection. *J Infect.* 2022;85(3):334–363. doi:10.1016/j.jinf.2022.05.028
 37. del Rio C, Malani PN. Update on the monkeypox outbreak. *JAMA.* 2022;328(10):921–922. doi:10.1001/jama.2022.14857
 38. World Health Organization. Vaccines and immunization for monkeypox: Interim guidance, 16 November 2022. Accessed November 16, 2022. <https://www.who.int/publications/i/item/WHO-MPX-Immunization>
 39. U.S. Food & Drug Administration. September 24, 2019 Approval Letter- JYNNEOS. Accessed September 3, 2023. <https://www.fda.gov/media/131079/download?attachment>
 40. U.S. Center for Disease Control and prevention. JYNNEOS Vaccine. Accessed August 22, 2023. <https://www.cdc.gov/poxvirus/mpox/interim-considerations/jynneos-vaccine.html>
 41. U.S. Center for Disease Control and prevention. Jynneos Vaccine Effectiveness. Accessed August 23, 2023. <https://www.cdc.gov/poxvirus/mpox/cases-data/JYNNEOS-vaccine-effectiveness.html>
 42. Bragazzi NL, Han Q, Iyaniwura SA, et al. Adaptive changes in sexual behavior in the high-risk population in response to human monkeypox transmission in Canada can help control the outbreak: insights from a two-group, two-route epidemic model. *J Med Virol.* 2023;95(4):e28575. doi:10.1002/jmv.28575
 43. Betancort-Plata C, Lopez-Delgado L, Jaén-Sánchez N, et al. Monkeypox and HIV in the Canary Islands: a different pattern in a mobile population. *Trop Med Infect Dis.* 2022;7(10):318. doi:10.3390/tropicalmed7100318
 44. Molero F, Silván-Ferrero P, Fuster-Ruiz de Apodaca MJ, Nouvilas-Pallejá E, Pérez-Garín D. Subtle and blatant perceived discrimination and well-being in lesbians and gay men in Spain: the role of social support. *Psicothema.* 2017;29(4):475–481. doi:10.7334/psicothema2016.296
 45. Guerras JM, Hoyos J, de la Fuente L, et al. Awareness and use of HIV Self-Testing among men who have sex with men remains low in Spain 2 years after its authorization. *Front Public Health.* 2022;10:888059. doi:10.3389/fpubh.2022.888059
 46. Liu MY, Li QH, Zhang YJ, et al. Spatial and temporal clustering analysis of tuberculosis in the mainland of China at the prefecture level, 2005–2015. *Infect Dis Poverty.* 2018;7(1):106. doi:10.1186/s40249-018-0490-8

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Shang W, Cao G, Wu Y, et al. Spatiotemporal cluster of mpox in men who have sex with men: a modeling study in 83 countries. *J Med Virol.* 2023;95:e29166. doi:10.1002/jmv.29166