BMC Research Notes

DATA NOTE Open Access

Cardiovascular health in perspective: a comprehensive five-year geodatabase of hospitalizations and environmental factors in Mashhad, Iran

Shahab MohammadEbrahimi[1*], Mohammad Dehghan[2] and Behzad Kiani[3]

Abstract

Objectives This data note presents a comprehensive geodatabase of cardiovascular disease (CVD) hospitalizations in Mashhad, Iran, alongside key environmental factors such as air pollutants, built environment indicators, green spaces, and urban density. Using a spatiotemporal dataset of over 52,000 hospitalized CVD patients collected over five years, the study supports approaches like advanced spatiotemporal modeling, artificial intelligence, and machine learning to predict high-risk CVD areas and guide public health interventions.

Data description This dataset includes detailed epidemiologic and geospatial information on CVD hospitalizations in Mashhad, Iran, from January 1, 2016, to December 31, 2020. It contains 52,176 confirmed CVD cases and includes demographic information such as age, gender, admission date, ICD-10 codes, occurrence of death, and length of hospital stay. The median age of patients was 64 years, with 54.44% male. A notable 9.41% of patients died during hospitalization. In addition to the CVD hospitalization case file and its shape file created by joining with 1301 census tracts, this dataset includes environmental factors such as air quality indicators (SO_2, $PM_{2·5}$, CO, etc.). It also incorporates socio-economic variables (population density, illiteracy, and unemployment rates), public infrastructure, and built environment data, providing a comprehensive view of cardiovascular health in Mashhad.

Keywords Cardiovascular disease (CVD), Space-time analysis, Geographical information system (GIS), Public health, Epidemiology, Iran

*Correspondence:
Shahab MohammadEbrahimi
ShahabOdd@gmail.com
[1]Department of Medical Informatics, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran
[2]Department of Anesthesiology and Critical Care, Tehran University of Medical Sciences, Tehran, Iran
[3]UQ Centre for Clinical Research, Faculty of Health Medicine and Behavioural Sciences, The University of Queensland, Brisbane, Australia

Objective

Cardiovascular disease (CVD) remains one of the leading causes of mortality worldwide, responsible for a considerable proportion of global deaths each year [1]. The growing burden of CVD highlights the need to understand not only its biological risk factors but also the broader environmental and socioeconomic determinants that contribute to its burden. Environmental factors, such as air pollution, socioeconomic conditions, and the built environment, are increasingly recognized for their role in exacerbating cardiovascular conditions [2, 3]. Urban environments pose distinct challenges, where factors such as high population density, increased exposure to pollutants, and restricted access to green spaces can amplify the risks associated with cardiovascular disease [4]. Thus, studying the spatial distribution of CVD in conjunction with these modifying factors is crucial for addressing this global health challenge.

This data note presents data on the geographic distribution of hospitalized CVD patients in Mashhad, Iran. It also outlines potential environmental variables, such as air quality, green spaces, and urban density, that may be associated with cardiovascular health outcomes. To achieve this, we systematically collected, integrated, and compiled a comprehensive geospatial and epidemiologic dataset of hospitalized CVD patients across a five-year period. By leveraging this extensive dataset, which offers high spatial and temporal granularity, researchers can conduct various analyses to uncover spatial patterns of CVD and explore the relationship between environmental stressors and cardiovascular risk [5, 6].

The dataset offers significant potential for research applications, including the identification and prediction of high-risk CVD areas through spatial epidemiology, artificial intelligence, and machine learning. Moreover, this geodatabase can significantly inform urban planning and environmental policy by providing insights into how urban design, land use, and transportation systems influence cardiovascular health. By identifying areas with high concentrations of pollutants or limited access to healthcare services, urban planners can develop targeted interventions to mitigate health risks. Additionally, the data can support the formulation of evidence-based policies aimed at improving public health outcomes, such as enhancing green spaces, optimizing public transportation routes, and implementing stricter environmental regulations.

Data description

Table 1 shows the details of the four data files plus the data dictionary. The dataset encompasses detailed geospatial and epidemiologic information on CVD hospitalizations in Mashhad, Iran, collected from January 1, 2016, to December 31, 2020. The data were sourced from the Hospital Information Systems (HIS) managed by Mashhad hospitals, which serves as a comprehensive repository of information for patients admitted to hospitals. The dataset includes hospitalizations for various cardiovascular disorders, classified using ICD-10 codes, including ischemic heart disease (I20-I25), atrial fibrillation (I48), heart failure (I50), ischemic stroke (I63, I65-I67, G45, G46), and atherosclerotic disease (I70-I77.1).

A series of preprocessing steps were implemented to refine the initial raw dataset. Address fields were manually corrected for spelling and completeness. Records were excluded based on specific criteria: patients residing outside Mashhad, those staying at hotels or inns due to the city's prominence as a pilgrimage site, and any records with missing address information. These exclusion criteria ensure that the dataset accurately reflects the local population and enhances the robustness of subsequent analyses. After preprocessing, the final dataset comprises 52,176 hospitalization records, including details such as age, sex, admission dates, ICD-10 codes, length of stay (LOS), and in-hospital mortality status (Data file 1). The dataset comprises 54.44% male patients, with a median age of 64 years (interquartile range (IQR): 53–76 years) across both sexes. The median LOS in the hospital was 2 days (IQR: 1–5 days). A notable 9.41% of the CVD patients succumbed to the disease, resulting in 4,907 deaths during the study period. While 2019 recorded the highest level of CVD admissions, Joinpoint Regression analysis revealed a significant shift in trends starting in late 2019, coinciding with the onset of COVID-19. From

Table 1 Overview of data files

Label	Name of data file	File types	Data repository and identifier
Data file 1	Mshd_CVD_Casefile	MS Excel (.xlsx)	Harvard Dataverse https://doi.org/10.7910/DVN/NLOJTJ [17]
Data file 2	Mshd_CVD_Geospatialfile	Shape file (.shp)	Harvard Dataverse https://doi.org/10.7910/DVN/NLOJTJ [17]
Data file 3	Mshd_Environmentalfactors	Shape file (.shp)	Harvard Dataverse https://doi.org/10.7910/DVN/NLOJTJ [17]
Data file 4	Data usage agreement	Portable Document Format (.pdf)	Harvard Dataverse https://doi.org/10.7910/DVN/NLOJTJ [17]
Data dictionary	Help file	MS Excel (.xlsx)	Harvard Dataverse https://doi.org/10.7910/DVN/NLOJTJ [17]

November 2019 to December 2020, CVD rates declined notably, with monthly percent changes (MPC) of -6.04 for males and −6.49 for females, likely due to reduced healthcare access and changes in health behaviors during the pandemic [7, 8].

To create a geospatial dataset, patient addresses were geocoded and mapped using the Google My Map service. To enhance confidentiality, a 100-meter jittering technique was employed, randomly displacing each geocoded point within a 100-meter radius to obscure exact locations while preserving spatial accuracy [9]. All geocoded cases were then spatially joined with the census tracts (n: 1301) for detailed geographic analysis. After excluding 44 cases outside the geographical boundaries, the final CVD geodatabase comprised 52,132 cases (Data file 2). This rigorous approach to geocoding enhances the dataset's robustness, ensuring reliable spatial analysis. Please refer to Supplementary Material 1 and 2 for heat plots on CVD hospitalizations, deaths, LOS, and the spatial distribution of Standardized Incidence Ratios (SIR) for CVD hospitalizations during the study period.

The study area spanned 1,301 census tracts, covering a total population of 3,001,225 individuals, based on the 2016 population and housing census (male: 50.11%). In addition to the CVD hospitalization data, this dataset also includes environmental factors relevant to cardiovascular health, which are provided as a geographical shape file (Data file 3). Rationale for selecting these environmental variables was grounded in previous literature that has examined the association between CVD and environmental factors [10–15]. Access to these variables from various sources was crucial for our analysis. These factors encompass air quality indicators, such as sulfur dioxide (SO_2), particulate matter ($PM_{2.5}$ and PM_{10}), carbon monoxide (CO), nitrogen dioxide (NO_2), and ozone (O_3) levels. The average concentration of air pollutants was measured using data collected from 22 monitoring stations across Mashhad. Census-level exposure was estimated using inverse distance weighting (IDW) interpolation, where pollutant values were calculated for grid rectangles based on the proximity to nearby monitoring stations [16]. Finally, overall pollutant averages for each census were derived from the interpolated grid values within each census tract. Public infrastructure data, including the distribution of transportation facilities such as bus stops, fuel stations, and road intersections, is included. Additionally, built environment data, incorporating the Normalized Difference Vegetation Index (NDVI), provides information on green spaces and densities of commercial and industrial centers. All data is sourced from the municipal council. Socio-economic variables, such as population density, illiteracy rates, and unemployment rates—derived from the 2016 population and housing census—are also incorporated to offer insights into the broader environmental and socio-economic context of cardiovascular disease. Notably, data files 2 and 3 can be joined using the "Tractid" feature. Data file 4 includes the data usage agreement that researchers must accept to access Data file 1. Table 1 presents a summary of the dataset, which has been made available on the Harvard Dataverse [17].

Limitations

- The use of hospital admission data inherently limits the scope of our dataset, as it may lead to an underestimation of the true incidence of CVDs. This limitation arises from the exclusion of patients treated in outpatient settings or those who did not seek medical care. Consequently, the results may not fully capture the impact of CVDs in the broader population, highlighting the need for caution when interpreting the results.
- Although we analyzed CVD trends over a five-year period, the study period overlaps with the COVID-19 pandemic, which likely influenced temporal patterns in CVD hospitalizations and outcomes. This overlap may affect the generalizability of the results to non-pandemic periods.
- Future research should expand to include data from outpatient and primary care facilities and assess the long-term effects of the COVID-19 pandemic on CVD trends in the region.

Abbreviations

CVD	Cardiovascular diseases
GIS	Geographical information system
ICD-10	International classification of diseases, 10_{th} revision
HIS	Hospital information system
SIR	Standardized incidence ratio
IQR	Inter-quartile range
SO_2	Sulfur dioxide
$PM_{2.5}$ & PM_{10}	Particulate matter
CO	Carbon monoxide
NO_2	Nitrogen dioxide
O_3	Ozone
IDW	Inverse distance weighting
NDVI	Normalized difference vegetation index

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13104-025-07087-5.

Supplementary material 1

Supplementary material 2

Acknowledgements

We would like to express our deepest gratitude to Mashhad University of Medical Sciences for funding this research and offering the data.

Author contributions

S.M. drafted the manuscript, while B.K. and M.D. provided critical revisions. M.D. also validated the CVD data. S.M. contributed to data cleaning,

preprocessing, and preparing the dataset for publication. B.K., as the research leader, secured the funding for this study. All authors read and approved the final manuscript.

Funding
This study was funded by Mashhad University of Medical Sciences (Fund Number: 990724).

Data availability
Data files 2 and 3 included in this data note are freely accessible on the Harvard Dataverse at https://doi.org/10.7910/DVN/NLOJTJ [10]. In contrast, Data File 1 contains indirect identification attributes of hospitalized CVD patients and is restricted to individuals who accept specific usage conditions outlined in Data File 4 (Data usage agreement) and have received ethical approval from an appropriate institution for their research. Furthermore, each request will undergo a thorough review, and only the essential variables will be shared to protect patient privacy. Please see Table 1 and the references for details and links to the data.

Declarations

Ethics approval and consent to participate
The study was approved by the ethical committee of Mashhad University of Medical Sciences with the reference number IR.MUMS.MEDICAL.REC.1399.422. As well as informed consent was not required to be obtained due to the nature of the study.

Consent for publication
All authors have reviewed and approved the manuscript for publication.

Competing interests
The authors declare no competing interests.

Received: 20 October 2024 / Accepted: 6 January 2025
Published online: 13 January 2025

References

1. Vivarelli S, Costa C, Teodoro M, Giambò F, Tsatsakis AM, Fenga C. Polyphenols: a route from bioavailability to bioactivity addressing potential health benefits to tackle human chronic diseases. Arch Toxicol. 2023;97(1):3–38.
2. Hajar R. Risk factors for coronary artery disease: historical perspectives. Heart Views off J Gulf Heart Assoc. 2017;18(3):109–14.
3. Wang C, Wang C, Liu M, Chen Z, Liu S. Temporal and spatial trends of ischemic heart disease burden in Chinese and subgroup populations from 1990 to 2016: socio-economical data from the 2016 global burden of disease study. BMC Cardiovasc Disord. 2020;20(1):243.
4. Roux AVD, Merkin SS, Arnett D, Chambless L, Massing M, Nieto FJ, et al. Neighborhood of Residence and incidence of Coronary Heart Disease. N Engl J Med. 2001;345(2):99–106.
5. Ghimire U, Yasmin S, Chand S, Timalsena BK, Bhat T, Thapa S, et al. Cardiovascular disease risk factors distribution and clustering across different geographic levels in Nepal. Am J Hum Biol. 2022;34(9):e23787.
6. Yan S, Liu G, Chen X. Spatiotemporal distribution characteristics and influencing factors of the rate of cardiovascular hospitalization in Ganzhou city of China. Front Cardiovasc Med. 2023;10:1225878.
7. Mansouritorghabeh H, Bagherimoghaddam A, Eslami S, Raouf-Rahmati A, Hamer DH, Kiani B, et al. Spatial epidemiology of COVID-19 infection through the first outbreak in the city of Mashhad, Iran. Spat Inf Res. 2022;30(5):585–95.
8. MohammadEbrahimi S, Mohammadi A, Bergquist R, Akbarian M, Arian M, Pishgar E, et al. A spatial-epidemiological dataset of subjects infected by SARS-CoV-2 during the first wave of the pandemic in Mashhad, second-most populous city in Iran. BMC Res Notes. 2021;14(1):292.
9. 'Unmasking' masked. Address data: a medoid geocoding solution. MethodsX. 2023;10:102090.
10. Wang B, Gu K, Dong D, Fang Y, Tang L. Analysis of spatial distribution of CVD and Multiple Environmental Factors in urban residents. Comput Intell Neurosci. 2022;2022(1):9799054.
11. Malambo P, Kengne AP, Villiers AD, Lambert EV, Puoane T. Built Environment, selected risk factors and Major Cardiovascular Disease outcomes: a systematic review. PLoS ONE. 2016;11(11):e0166846.
12. Bhatnagar A. Environmental determinants of Cardiovascular Disease. Circ Res. 2017;121(2):162–80.
13. Münzel T, Hahad O, Sørensen M, Lelieveld J, Duerr GD, Nieuwenhuijsen M, et al. Environmental risk factors and cardiovascular diseases: a comprehensive expert review. Cardiovasc Res. 2021;118(14):2880–902.
14. Hadianfar A, Küchenhoff H, MohammadEbrahimi S, Saki A. A novel spatial heteroscedastic generalized additive distributed lag model for the spatio-temporal relation between PM2.5and cardiovascular hospitalization. Sci Rep. 2024;14(1):1–12.
15. Sagheer U, Al -Kindi, Sadeer, Abohashem S, Phillips CT, Rana JS, Bhatnagar A, et al. Environmental Pollution and Cardiovascular Disease. JACC Adv. 2024;3(2):100805.
16. Wong DW, Yuan L, Perlin SA. Comparison of spatial interpolation methods for the estimation of air quality data. J Expo Sci Environ Epidemiol. 2004;14(5):404–15.
17. Mohammad Ebrahimi S. Cardiovascular Hospitalizations in Mashhad, Iran, Over a Five-Year Period: A Geodatabase. Harvard Dataverse 2024. Available from: https://doi.org/10.7910/DVN/NLOJTJ

Publisher's note