Perspectives

State of the Art of Lifecourse Cohort Establishment

Shaoqing Dai[1,2]; Ge Qiu[1,2]; Yuchen Li[2,3,4]; Shuhan Yang[1,2]; Shujuan Yang[2,5]; Peng Jia[1,2,6,7,8,#]

The global rise in non-communicable diseases (NCDs) presents significant public health challenges. Effectively managing and preventing NCDs necessitates a thorough understanding of their causes and progression, which can be achieved through a lifecourse approach to determine past exposures' impact before NCD onset. However, this approach requires robust backing from data, specifically lifecourse cohort data, which are generally insufficient. To overcome this obstacle, three primary strategies have been employed to establish such cohorts: active follow-up cohorts, registry-based datasets, and technology-based data collection and simulation methods.

ACTIVE FOLLOW-UP COHORTS

Continuous health and behavior monitoring in active follow-up cohorts is essential for early identification and management of risk factors. Despite being resource-intensive, collaboration among global epidemiologists has facilitated access to extensive long-term follow-up cohorts, enhancing lifecourse epidemiological research capabilities.

The UK Biobank is an exemplary population-based prospective cohort study focused on the genetic and non-genetic factors influencing diseases in adults and the elderly (*1*). It aimed for an extensive evaluation of exposures, meticulous follow-up, and detailed characterization of various health outcomes. By 2010, the UK Biobank had recruited 500,000 participants ranging from 40 to 69 years old, amassing an extraordinary array of baseline data and biological samples. Subsequently, three follow-up surveys were conducted in 2012–2013, 2014, and 2019. The database has been continually enhanced since 2012 with additional types of data, including monthly blood samples and nuclear magnetic resonance spectroscopy data, among others. To date, the UK Biobank has compiled a comprehensive dataset featuring details on over 8,500 deaths, upward of 75,000 cancer cases, and more than 600,000 hospital admissions.

The China Kadoorie Biobank (CKB) has made significant progress (*2*). The baseline survey, conducted from 2004 to 2008, covered 10 specific regions and included questionnaire data, physical measurements, and blood samples. In 2013–2014, a second survey was conducted with 25,091 participants aged 30–79 years, followed by a third survey in 2020–2021 with 25,087 participants (*3*). Importantly, a substantial cohort of over 22,000 individuals participated in at least two follow-ups, forming a crucial basis for future longitudinal analyses. The availability of multiple waves of data collected at different time points will enable detailed investigations into the trends of risk factors related to major diseases.

Cohort studies that integrate the lifecourse perspective have significantly enhanced our comprehension of the ramifications of exposures during the early stages of life. The Human Early-Life Exposome (HELIX) project exemplifies such research endeavors, focusing on delineating the spectrum of environmental exposures during prenatal and early childhood phases. The project investigates the associations between these exposures and critical pediatric health outcomes, such as growth patterns, obesity prevalence, neurodevelopmental progress, and respiratory health. To achieve its aims, the HELIX project utilizes an array of investigative tools, including biomarker assessments, omics technologies, geospatial analyses, monitoring through wearable devices, and sophisticated statistical methodologies (*4*). Operating within the framework of a "lifecourse exposome" model, HELIX aggregates data from six established birth cohort studies spanning various regions in Europe. It is undertaking the development of comprehensive exposure models for its entire cohort, which embraces over 32,000 mother-child pairs, specifically concentrating on children within the 6–11 year age bracket.

Cohort studies focusing on specific NCDs have been instrumental in advancing lifecourse epidemiology. A notable example is the Framingham Heart Study (FHS), a pioneering intergenerational longitudinal study that began in 1948 with the goal of enhancing our understanding of cardiovascular disease

epidemiology in the United States (*5*–*6*). To date, the FHS has followed up with a total of 15,447 participants, with more than 9,000 followed until death as of 2019. The study population encompasses six cohorts: the Original Cohort (*n*=5,209, ages 28–74 in 1948), Offspring Cohort (*n*=5,124, ages 5–70 in 1971), Omni Generation 1 Cohort (*n*=506, ages 27–78 in 1994), Third Generation Cohort (*n*=4,095, ages 19–72 in 2002), New Offspring Spouse Cohort (*n*=103, ages 47–85 in 2003), and Omni Generation 2 Cohort (*n*=410, ages 20–80 in 2003). The study is known for its detailed participant characterization, regular follow-up examinations, and comprehensive surveillance of both cardiovascular and non-cardiovascular endpoints, providing a solid basis for research into various health outcomes.

The establishment of cohorts focused on detailed occupational experiences is a growing trend in lifecourse epidemiology. A preeminent example is the Nurses' Health Study, a comprehensive prospective cohort that investigates risk factors for major chronic diseases in women (*7*). Initiated in 1976, the Nurses' Health Study is an ongoing project that now includes both male and female nurses. This project has enrolled more than 275,000 participants across three generations: the inaugural cohort from 1976 (ages 30–55), the second cohort from 1989 (ages 25–42), and the third cohort from 2010 (ages 19–46). It conducts active follow-up and gathers lifestyle data every four years. Additionally, the study has collected blood samples and DNA from buccal cell extractions. This cohort also integrates tumor sample information from participants who are part of other databases, thereby providing foundational data for more comprehensive analysis.

REGISTRY-BASED DATASETS

The utilization of lifecourse cohort studies leverages readily accessible information from various governmental databases, including those related to residency, education, housing, taxes, driver's licenses, insurance, and medical records. While this method may not always capture specialized data such as behavioral and psychological factors necessary for rigorous epidemiological investigations, it is a cost-effective alternative to the creation and maintenance of active follow-up cohorts over the long term. As a feasible method for building lifecourse cohorts in the present context, it offers a practical solution when compared to other methods. In certain European nations, merging residency with medical records facilitates the efficient collection of baseline demographic and health information from broad administrative registries. This process simplifies the establishment of cohorts that encompass extensive time periods. For instance, the Nordic countries — comprising Denmark, Finland, Norway, and Sweden — operate comprehensive registries that cover all citizens, enabled by unique personal identification numbers which allow for the cross-referencing of multiple information systems. By interconnecting various medical record databases, which include data from the Danish healthcare system, the Swedish Medical Birth Registries, the Nordic Cancer Registries, Prescription Registries, Medical Birth Registries, and Patient Registries, with residency records, these countries have created registry-based datasets that span almost 40 years. Such datasets are highly beneficial for diverse health-related research projects. For example, these datasets have been used to study the association between adult stress and atrial fibrillation risk (*8*), assess the real-world effectiveness of medications like liraglutide in the clinical management of cardiovascular diseases (*9*), and probe the potential link between prenatal antibiotic exposure and childhood leukemia incidence (*10*). Another exemplary registry-based dataset is within the UK's National Health Service. Utilizing medical record data from national registers that cover patients who were hospitalized for their first acute ischemic stroke or primary intracerebral hemorrhage in England from 2013 to 2016, which included a total of 145,324 individuals, studies have investigated socioeconomic differences in initial stroke hospitalization rates, evaluated care quality, and assessed post-stroke survival rates among the adult populace in England (*11*). In Australia, a cohort of 85,547 individuals was developed by integrating data from the National Diabetes Services Scheme — which supports patients by providing diabetes-related products at subsidized rates and disseminating essential information — and the National Death Index to track mortality rates among Australians diagnosed with type 1 diabetes (*12*).

Governmental resources beyond healthcare, such as those related to insurance, education, and taxation, are increasingly recognized as valuable for constructing registry-based datasets to tackle multifaceted issues in lifecourse epidemiology. For instance, in Sweden, the amalgamation of longitudinal health insurance and labor market data with registry information about the resident population produced a comprehensive dataset

covering 1990–2007, which included over 6 million individuals (6.04 million). This extensive dataset was leveraged to explore the association between individual socioeconomic factors — insurance, education, taxation — and mortality rates throughout the adults' life span (*13*). In Norway, a distinct registry-based dataset was assembled, containing data on 3.1 million individuals aged 18–69. It integrated information from the national road accident registry with the Norwegian prescription database to assess the risk of road traffic accidents in relation to prescription medication usage among drivers (*14*). Most registry-based datasets are constructed with the resident population as a foundation, linked to medical records, and further enriched by integrating additional governmental resources.

TECHNOLOGY-BASED DATA COLLECTION AND SIMULATION METHODS

To enhance the caliber of existing cohorts beyond the capabilities of traditional survey methods and linkages, it is imperative to incorporate technology-based data collection and simulation techniques. Such methods leverage sophisticated, interactive devices to gather real-time, uninterrupted data that are more detailed in both spatial and temporal aspects compared to conventional epidemiological data collection. For instance, advancements in internet communication technology have underscored the growing relevance of technology-based data collection simulations in developing lifecourse cohorts. A notable example is the UK Biobank initiative, where Axivity AX3 tri-axial wrist physical activity monitors were distributed to 100,000 participants, capturing high-frequency (100 Hz) triaxial acceleration over a week. This yielded a pivotal dataset for an in-depth analysis of daily physical activities and exposure to real-world environments (*15*). Additionally, cutting-edge fiber technology has facilitated the integration of wearable devices into clothing, conveniently tracking behaviors and health status. A recent study introduced a mechanical design for semiconductor fibers, functioning as sensors, actuators, energy harvesters and storages, displays, and healthcare devices (*16*).

Environmental factors persistently influence both individual behaviors and health outcomes. They can be comprehensively monitored using remote sensing technology, utilizing sensors aboard satellites for broad environmental surveillance globally. Direct measurement or straightforward calculation of certain environmental variables is possible using spectral information obtained from these sensors (*17–19*). For instance, airborne sensors on aircraft and unmanned aerial vehicles, such as drones, can directly capture urban built environment features including building outlines, road widths, and traffic density (*20–21*). Vegetation coverage, indicated by parameters like greenness from trees and grasslands, can be quantified through spectral data collected by both airborne sensors and high-resolution satellites (*22–23*). Additionally, certain environmental factors require more complex algorithms and supplementary data for accurate retrieval. For example, the concentrations of fine particulate matter with a diameter of ≤2.5 μm ($PM_{2.5}$), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and ozone (O_3), as well as the chemical compositions of $PM_{2.5}$, at different temporal resolutions (e.g., daily, monthly) can be derived through a combination of satellite-derived, ground-based monitoring, and other auxiliary data (e.g. meteorological and land use data) (*24*).

FUTURE PERSPECTIVES FOR LIFECOURSE COHORT DEVELOPMENT

Each of the three discussed methodologies has its advantages and disadvantages. Active follow-up cohort studies are highly effective for exploring specific public health concerns, yet they are limited by significant time and financial demands, and they do not provide a comprehensive perspective on overall human health. Registry-based datasets, by contrast, offer greater cost efficiency due to pre-existing governmental funding, eliminating the need for additional cohort establishment. However, their reliance on medical and death records means they might not fully capture the entire scope of health issues throughout an individual's life, potentially limiting their ability to offer a complete picture of public health trends. While innovative, technology-based data collection and simulation methods have the potential to fill many of the voids inherent in traditional models, the interdisciplinary nature of these new approaches poses challenges to conventional sectors and professionals, necessitating increased cross-disciplinary collaboration to be successfully implemented.

The advancement of lifecourse cohort studies should

consider three key aspects, informed by current strategic efforts. Firstly, while active follow-up cohorts typically emphasize the health of adults due to the higher incidence of NCDs in this group, there is a need to shift the baseline of future cohorts to earlier life stages, such as initiating at birth. This approach will enable the tracking of health trajectories from infancy, through positive child development studies (*22*), and across the entire lifespan, allowing for a more comprehensive understanding of health evolution. Secondly, the scope of existing registry-based datasets is often limited in the variety and quantity of data they encompass, and frequently fail to integrate medical records. Consequently, it is essential to establish (real-time) data platforms that can amalgamate diverse information sources while rigorously safeguarding data confidentiality. Thirdly, the current application of cutting-edge technologies in the collection and simulation of technology-based data is restricted. The field of spatial lifecourse health, which has evolved from spatial lifecourse epidemiology (*23*) and lies at the confluence of spatial science and public health, leverages sophisticated technologies and methodologies. These include geoinformatics, remote sensing, global navigation satellite systems, the Internet of Things, artificial intelligence, mathematical statistics, bioinformatics, systems science, data science, and augmented, virtual, and mixed reality. These tools enable highly precise assessments of environmental, behavioral, psychological, physiological, and biological risk factors affecting health, while also examining their long-term impacts and underlying causal mechanisms. This domain has significantly contributed to the research of NCDs, infectious diseases, public health monitoring, and 'one health', collecting diverse sets of data in novel ways that address significant data deficiencies present in traditional fields and sectors (*24*). Therefore, it possesses considerable potential to drive the integration of efforts in establishing authentic lifecourse cohort studies.

Conflicts of interest: No conflicts of interest.

Funding: Supported by the National Natural Science Foundation of China (42271433), the National Key R&D Program of China (2023YFC3604701), the "0 to 1" Innovation Research Project of Sichuan University (2023CX21), the Key R&D Project of Sichuan Province (2023YFS0251), Renmin Hospital of Wuhan University (JCRCYG-2022-003), the Wuhan University Specific Fund for Major School-level Internationalization Initiatives (WHU-GJZDZX-PT07), and the International Institute of Spatial Lifecourse Health (ISLE).

doi: 10.46234/ccdcw2024.058

[#] Corresponding author: Peng Jia, jiapeng@hotmail.com.

[1] School of Resource and Environmental Sciences, Wuhan University, Wuhan City, Hubei Province, China; [2] International Institute of Spatial Lifecourse Health (ISLE), Wuhan University, Wuhan City, Hubei Province, China; [3] MRC Epidemiology Unit, University of Cambridge, Cambridge, UK; [4] Department of Geography, The Ohio State University, Columbus, OH, USA; [5] West China School of Public Health and West China Fourth Hospital, Sichuan University, Chengdu City, Sichuan Province, China; [6] Hubei Luojia Laboratory, Wuhan City, Hubei Province, China; [7] School of Public Health, Wuhan University, Wuhan City, Hubei Province, China; [8] Renmin Hospital, Wuhan University, Wuhan City, Hubei Province, China.

Submitted: February 26, 2024; Accepted: March 08, 2024

REFERENCES

1. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 2015;12(3):e1001779. https://doi.org/10.1371/journal.pmed.1001779.
2. Chen Z, Chen J, Collins R, Guo Y, Peto R, Wu F, et al. China Kadoorie Biobank of 0. 5 million people: survey methods, baseline characteristics and long-term follow-up. Int J Epidemiol 2011;40(6):1652 – 66. https://doi.org/10.1093/ije/dyr120.
3. Walters RG, Millwood IY, Lin K, Schmidt Valle D, McDonnell P, Hacker A, et al. Genotyping and population characteristics of the China Kadoorie Biobank. Cell Genom 2023;3(8):100361. https://doi.org/10.1016/j.xgen.2023.100361.
4. Vrijheid M, Slama R, Robinson O, Chatzi L, Coen M, van den Hazel P, et al. The human early-life exposome (HELIX): project rationale and design. Environ Health Perspect 2014;122(6):535 – 44. https://doi.org/10.1289/ehp.1307204.
5. Andersson C, Johnson AD, Benjamin EJ, Levy D, Vasan RS. 70-year legacy of the Framingham heart study. Nat Rev Cardiol 2019;16(11):687 – 98. https://doi.org/10.1038/s41569-019-0202-5.
6. Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham heart study and the epidemiology of cardiovascular disease: a historical perspective. Lancet 2014;383(9921):999 – 1008. https://doi.org/10.1016/S0140-6736(13)61752-3.
7. Colditz GA, Hankinson SE. The nurses' health study: lifestyle and health among women. Nat Rev Cancer 2005;5(5):388 – 96. https://doi.org/10.1038/nrc1608.
8. Chen H, Janszky I, Rostila M, Wei D, Yang F, Li J, et al. Bereavement in childhood and young adulthood and the risk of atrial fibrillation: a population-based cohort study from Denmark and Sweden. BMC Med 2023;21(1):8. https://doi.org/10.1186/s12916-022-02707-4.
9. Svanström H, Ueda P, Melbye M, Eliasson B, Svensson AM, Franzén S, et al. Use of liraglutide and risk of major cardiovascular events: a register-based cohort study in Denmark and Sweden. Lancet Diabetes Endocrinol 2019;7(2):106 – 14. https://doi.org/10.1016/S2213-8587(18)30320-6.
10. Hjorth S, Pottegård A, Broe A, Hemmingsen CH, Leinonen MK, Hargreave M, et al. Prenatal exposure to nitrofurantoin and risk of childhood leukaemia: a registry-based cohort study in four Nordic countries. Int J Epidemiol 2022;51(3):778 – 88. https://doi.org/10.1093/ije/dyab219.
11. Bray BD, Paley L, Hoffman A, James M, Gompertz P, Wolfe CDA, et al. Socioeconomic disparities in first stroke incidence, quality of care, and survival: a nationwide registry-based cohort study of 44 million adults in England. Lancet Public Health 2018;3(4):e185 – 93. https://doi.org/10.1016/S2468-2667(18)30030-6.

12. Huo LL, Harding JL, Peeters A, Shaw JE, Magliano DJ. Life expectancy of type 1 diabetic patients during 1997–2010: a national Australian registry-based cohort study. Diabetologia 2016;59(6):1177 – 85. https://doi.org/10.1007/s00125-015-3857-4.
13. Katikireddi SV, Niedzwiedz CL, Dundas R, Kondo N, Leyland AH, Rostila M. Inequalities in all-cause and cause-specific mortality across the life course by wealth and income in Sweden: a register-based cohort study. Int J Epidemiol 2020;49(3):917 – 25. https://doi.org/10.1093/ije/dyaa053.
14. Engeland A, Skurtveit S, Mørland J. Risk of road traffic accidents associated with the prescription of drugs: a registry-based cohort study. Ann Epidemiol 2007;17(8):597 – 602. https://doi.org/10.1016/j.annepidem.2007.03.009.
15. Khurshid S, Weng LC, Al-Alusi MA, Halford JL, Haimovich JS, Benjamin EJ, et al. Accelerometer-derived physical activity and risk of atrial fibrillation. Eur Heart J 2021;42(25):2472 – 83. https://doi.org/10.1093/eurheartj/ehab250.
16. Wang ZX, Wang Z, Li D, Yang CL, Zhang QC, Chen M, et al. High-quality semiconductor fibres via mechanical design. Nature 2024;626(7997):72 – 8. https://doi.org/10.1038/s41586-023-06946-0.
17. Jia P, Stein A. Using remote sensing technology to measure environmental determinants of non-communicable diseases. Int J Epidemiol 2017;46(4):1343 – 4. https://doi.org/10.1093/ije/dyw365.
18. Mei K, Huang H, Xia F, Hong A, Chen X, Zhang C, et al. State-of-the-art of measures of the obesogenic environment for children. Obes Rev 2021;22(S1):e13093. https://doi.org/10.1111/obr.13093.
19. Wang QJ, Duoji Z, Feng CT, Fei T, Ma H, Wang SM, et al. Associations and pathways between residential greenness and hyperuricemia among adults in rural and urban China. Environ Res 2022;215:114406. https://doi.org/10.1016/j.envres.2022.114406.
20. Yu WQ, Liu Z, La Y, Feng CT, Yu B, Wang QJ, et al. Associations between residential greenness and the predicted 10-year risk for atherosclerosis cardiovascular disease among Chinese adults. Sci Total Environ 2023;868:161643. https://doi.org/10.1016/j.scitotenv.2023.161643.
21. Yang SJ, Feng CT, Fei T, Wu D, Feng L, Yuan FS, et al. Mortality risk of people living with HIV under hypothetical intervention scenarios of $PM_{2.5}$ and HIV severity: a prospective cohort study. Sci Total Environ 2024;916:169938. https://doi.org/10.1016/j.scitotenv.2024.169938.
22. Zhao L, Shek DTL, Zou K, Lei YL, Jia P. Cohort profile: Chengdu positive child development (CPCD) survey. Int J Epidemiol 2022;51(3):e95 – 107. https://doi.org/10.1093/ije/dyab237.
23. Jia P. Spatial lifecourse epidemiology. Lancet Planet Health 2019;3(2):e57 – 9. https://doi.org/10.1016/S2542-5196(18)30245-6.
24. Jia P, Liu SY, Yang SJ. Innovations in public health surveillance for emerging infections. Annu Rev Public Health 2023;44:55 – 74. https://doi.org/10.1146/annurev-publhealth-051920-093141.