Large population-based cohorts: Essential platforms for medical research

Ying Jiang,[1,2,3] Shankuan Zhu,[1,2,3] and Jinling Tang[4,5,1,*]

[1]Department of Nutrition and Food Hygiene, Children's Hospital, Zhejiang University School of Medicine, National Clinical Research Center for Children and Adolescents' Health and Disease, Hangzhou 310058, China
[2]Chronic Disease Research Institute, School of Public Health, School of Medicine, Zhejiang University, Hangzhou 310058, China
[3]Binjiang Institute of Zhejiang University, Hangzhou 310000, China
[4]Faculty of Computer Science and Control Engineering, Shenzhen University of Advanced Technology, Shenzhen 518107, China
[5]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518000, China
*Correspondence: tangjinling@suat-sz.edu.cn

Received: October 13, 2025; Accepted: December 12, 2025; Published Online: December 13, 2025; https://doi.org/10.59717/j.xinn-med.2026.100178

Citation: Jiang Y., Zhu S. and Tang J. (2026). Large population-based cohorts: Essential platforms for medical research. *The Innovation Medicine* **4**:100178.

INTRODUCTION

Medical research can be broadly divided into basic research and population research. Basic research is conducted in laboratories on molecules, cells and animals aiming to dissect mechanisms and generate new theories. In contrast, population research is conducted in human populations to validate and assess the practical usefulness of basic research discoveries and thus forms the last gateway from theories to practice, although it also can generate theories.

Basic research and population research are, however, interdependent of and complementary to each other. Medical advances require balanced development between them. Over 50% of human health is determined by social, environmental and lifestyle factors, which can only be addressed through population-based research approaches. This further underscores the need for a strong population research infrastructure.

However, population research is generally less recognized in most countries, including the US.[1] Taking China as an example, China has already set up many laboratories for basic research, notably state and key laboratories,[2] which serve as national platforms for basic biomedical research. However, so far it is still unclear how a state population research platform should be constructed. No matter how it is conceived, a large population-based cohort will be an indispensable part of it.

Steps	Tasks	Major Principles/Precautions	Examples
1	Specify the nature of the cohort	Depending on the theme of research; cohorts normally do not need to set specific research questions	Natural cohort, birth cohort, disease cohort, drug cohort
2	Recruit cohort participants	The included should represent the sampling population	Random sampling is ideal but rarely practical
3	Collect baseline data and biological samples; testing of biological samples	Data collected should be as wide as possible around the major research theme, and should also be complete, high quality and standardized	Questionnaires and blood samples
4	Recollect at predetermined time points data on selected or all baseline variables; testing of biological samples	The same as above, minimize refusals and losses to follow-up	The same as above
	Follow up and collect continuously data on outcomes	The same as above, minimize losses to follow-up	The same as above
5	Share data among researchers	Ensure data security and privacy protection	Any researcher who conduct biomedical research in human populations
	Conducting studies by using cohort data	Follow the general principles and methods of epidemiological studies	Prospective study, case-control study, cross-sectional study

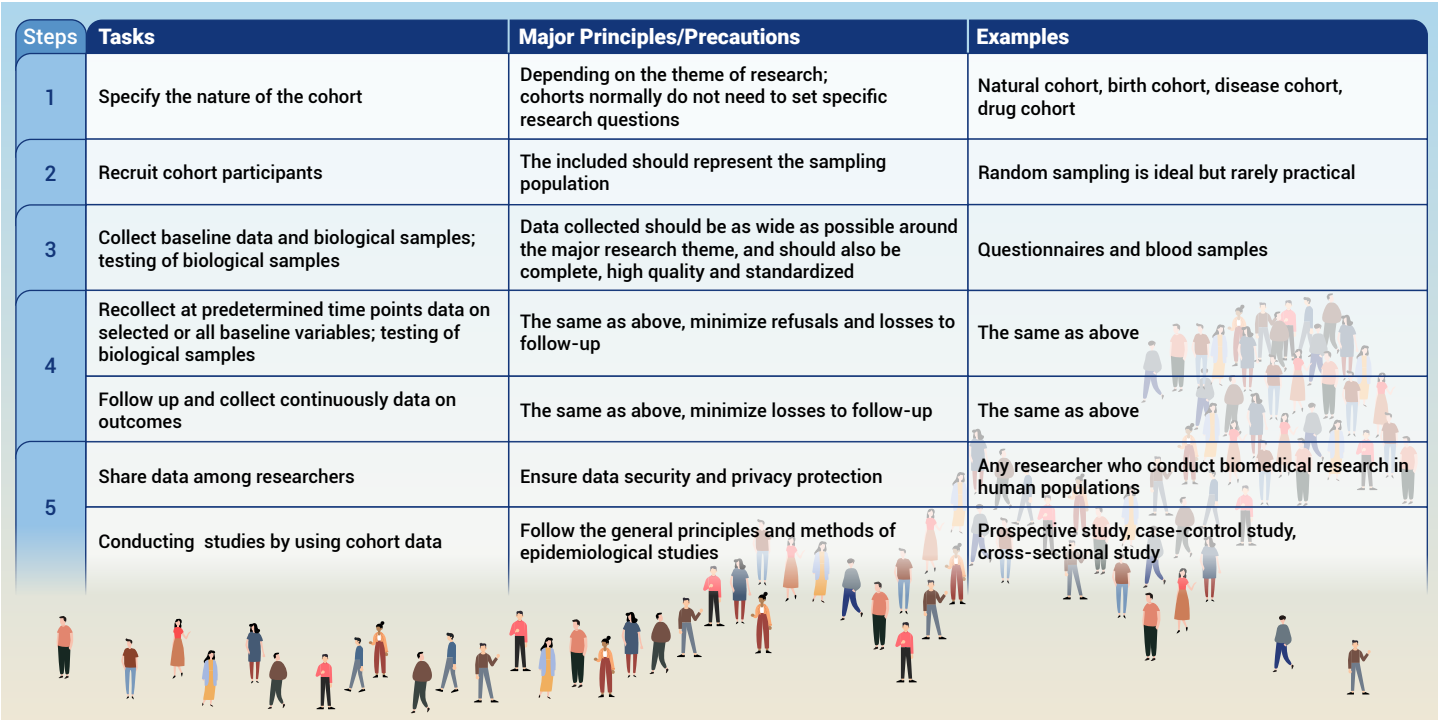

Figure 1. Building and using of large population-based cohorts for biomedical research: steps, tasks, principles and examples.

COHORTS AS POPULATION RESEARCH PLATFORMS IN MEDICINE

A cohort is a group of people with a common defining characteristic, from whom data can be collected at baseline and during follow-up for scientific research. In the past, cohorts were used by epidemiologists mostly to establish prospective studies identifying disease risk factors. For example, the Framingham Heart Study, which started in 1948, discovered the risk factors of cardiovascular disease. Importantly, the data from this study were later widely used for the investigation of many other research questions, marking the beginning of cohorts being used as research platforms for multiple studies.

Furthermore, with the rise of various omics technologies and the availability of routinely collected clinical data, the types of data a cohort can collect today have substantively expanded. These data can be used to investigate a much wider spectrum of research questions, including biological mechanisms in humans as well as clinical and public health practices.

It is now clear that large population-based cohorts are important platforms for population research, analogous to state laboratories for basic research. For example, the UK Biobank, a cohort of half a million people, in which £537.8 million have been invested so far, is an international population research platform.[3] In 2024 alone, over twenty thousand research teams from all over the world used the UK Biobank data for research, and it has produced

over 14,000 peer-reviewed publications since its onset.[4] In terms of both resources invested and the number of scientific investigations supported, the UK Biobank is comparable to a state laboratory in China, demonstrating the strategic role of large cohorts in biomedical research.

Since the late 20th century, scientists in such fields as computer science, omics research, wearables, and artificial intelligence have recognized the value of cohorts and actively participated in their establishment and utilization. Computers enable storage and analysis of large volumes of data; omics technologies provide new measurable components in biological samples; wearables allow continuous measurements, adding a chronological dimension; and artificial intelligence provides powerful new analytical tools, in particular for genetic, imaging, and multi-modal data. These developments greatly contribute to building and leveraging large cohorts for scientific research and open new venues for addressing health challenges.

PRINCIPLES AND PRECAUTIONS FOR BUILDING LARGE COHORTS

Regardless of emerging technologies, the rationale and principles in participant recruitment, data collection, bias control, data analysis and results interpretation[5] which have been well established over the past century by epidemiologists and medical statisticians, remain largely unchanged and must be observed in building and using cohorts for medical research (Figure 1).

First, participants recruited in a cohort should have a similar defining characteristic, for example, individuals from the general population in a population-based cohort, newborns in a birth cohort, and patients with the same disease in a disease-specific cohort. Cohorts further require that participants represent the population from which they are recruited, regardless of the defining characteristic. Losses to follow-up should be minimized to ensure outcome data are available for as many participants as possible.

Second, as previously argued, a cohort is a research platform that can be used for multiple investigations. Unlike traditional cohort studies, cohorts, in particular large ones, normally do not have to specify particular research questions, although they may focus on a broad research area. The usefulness of a cohort is determined by the width of the spectrum of information it collects: the more types of variables it includes, the more research questions it will be able to address. In data collection, although such new biomedical technologies as imaging, omics and wearables are important, they should not overshadow the behavioral, socioeconomic, and environmental factors.

Third, a cohort normally collects data on each variable for all individuals. If data on some variables are available for only a fraction of participants, the size of a study involving these variables will be limited by the variable with data available for the fewest participants. Besides completeness, data should also be collected with sufficient precision and accuracy, following consistent standards across all participants and throughout the observation period.

Although new methods can be used to build cohorts to save costs, for example, by using electronic medical records or merging existing cohorts, it should be noted that routine clinical data collected for patient care rather than research may not satisfy research's requirements for completeness, quality and consistency, and a consortium of existing cohorts with different purposes, standards and variable coverage does not form a single large cohort.

Fourth, how large should a cohort be? Scientifically, there is no clear consensus on this question. Currently, the largest cohorts of general populations are as large as half a million or even a million. Although these numbers are not a scientific estimate but a convenient figure, cohorts larger than those are not recommended. In fact, tens of thousands are sufficient for most research questions. First, past cohort studies and clinical trials have shown that a few thousand, or at most tens of thousands, of people are sufficient for identifying, with an acceptable statistical precision, a moderate effect that is useful for patient care. Second, building an extremely large cohort of general population to investigate rare diseases or exposures, such as rare genetic variants and gene-environment interactions, is not cost-effective; in such cases, case-control studies or cohorts of special populations will be more realistic. Third, recruiting a million people is already difficult but collection of some data, for example, whole genome sequencing, MRI and PET scan, and follow-up for decades, are many times more costly and/or time-consuming. Given limited resources, the number of variables and quality of data collected should thus not be compromised by an unnecessarily ambitious sample size.

Last but not least, similar to a research laboratory, the research potential of a cohort also depends on how widely its data are shared among researchers: the more widely the data are shared, the more achievements the cohort will yield. Thus, cohort funders should emphasize at the very beginning that the data of a cohort be widely shared among researchers. In addition to data sharing, the cohort participants can also be shared or used by other researchers to base on them to set up sub-cohorts.

LARGE COHORTS SHOULD BE BUILT AS NATIONAL POPULATION RESEARCH PLATFORMS

To strengthen population-based research, the UK and US have independently set up their own large cohorts. This suggests that China also needs to establish a large population cohort to support research to address its own healthcare issues, while also contributing data and research on genetic and environmental diversity to international efforts. Importantly, such large cohorts should be built as we build, support and use state laboratories in basic biomedical research.

In the meantime, cohort builders should also ensure stable financial supports, create sustainable running mechanisms, focus on common chronic diseases, collect information and biological samples as broadly as possible, ensure the quality, completeness, security and privacy protection of data, and finally share the data among researchers. Importantly, a large cohort, being a state strategic research investment, requires considerable and sustained financial support for it to succeed. The government certainly plays an important role in it, but other sources of support, such as the industry and charity, must also be eagerly and effectively solicited. If successfully established and widely shared, such a cohort will become a population research platform as important as a state laboratory for basic research.

Having said that, we must add that large cohorts also have limitations which are mostly determined by their representativeness of participants recruited, spectrum, completeness and quality of data collected, completeness of follow-up, and the extent to which their data are shared. In addition, large population-based cohorts do not eliminate the need for cohorts of special groups such as birth cohorts, drug cohorts and disease cohorts which can answer very different research questions.

CONCLUSION

Large population-based cohorts are so important for biomedical research that they should be invested in, established, and used in the way that strategic constructions for basic research are.

REFERENCES

1. Rosenberg R.N. (2003). Translating biomedical research to the bedside: A national crisis and a call to action. *JAMA* **289**:1305–1306. DOI:10.1001/jama.289.10.1305
2. Weinstein E.S., Lee C.N., Fedasiuk R., et al. (2022). China's state key laboratory system: A view into China's innovation system. *CSET* 1-27. DOI: 10.51593/20210019.
3. UK Biobank Homepage. https://www.ukbiobank.ac.uk/UK Biobank.
4. UK Biobank. (2025). UK Biobank's Wrap Up 2024. https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/news/uk-biobank-wrap-up-2024.
5. Hulley S.B., Cummings S.R., Browner W.S., et al. (2022). Designing clinical research (4th ed), Lippincott Williams & Wilkins. https://tilda.tcd.ie/epidemiology-biostatistics-course/course-material/assets/Class2/Designingclinicalresearch_4th-edition.pdf.

FUNDING AND ACKNOWLEDGMENTS

Ying Jiang and Shankuan Zhu were supported by National Key Research and Development Program of China (Grant No.: 2022YFC2705300) and Jinling Tang by the Shenzhen Science and Technology Program (Grant No.: KQTD20190929172835662) and Shenzhen Medical Research Funds (Grant No.: C2303001). The sponsors have no role in study design, data collection, data analysis, and results interpretation.

AUTHOR CONTRIBUTIONS

Professor Jinling Tang conceived the original ideas and revised the manuscript, Ms Ying Jiang integrated the materials, prepared the first draft and helped with revisions, and Professor Shankuan Zhu made important comments and suggestions to the improvement of the manuscript.

DECLARATION OF INTERESTS

Jinling Tang is an Editorial Board member of The Innovation Medicine and was blinded from reviewing or making final decisions on the manuscript. Peer review was handled independently of this member and their research group. The other authors declare no conflicts of interest.