COMMENTARY

Building and sharing medical cohorts for research

Guo-Bo Chen,[1,2,*] Siyang Liu,[3] Lei Zhang,[4] Tao Huang,[5] Xiaohua Tang,[1,6] Yixue Li,[7] and Changqing Zeng[8]

[1]Department of Genetic and Genomic Medicine, Zhejiang Provincial People's Hospital, Affiliated People's Hospital of Hangzhou Medical College, Hangzhou 310014, China
[2]Ningbo Institute of Life and Health Industry, University of Chinese Academy of Sciences, Ningbo 315000, China
[3]School of Public Health (Shenzhen), Shenzhen Campus of Sun Yat-sen University, Shenzhen 518017, China
[4]China National GeneBank, Shenzhen 518116, China
[5]Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai 200031, China
[6]Wenzhou Central Hospital, Dingli Clinical Medical School of Wenzhou Medical University, Wenzhou 325000, China
[7]Guangzhou Laboratory, Guangzhou 510320, China
[8]Henan Academy of Sciences, Zhengzhou 450046, China
*Correspondence: chenguobo@gmail.com

Received: January 17, 2024; Accepted: April 1, 2024; Published Online: April 2, 2024; https://doi.org/10.1016/j.xinn.2024.100623

Citation: Chen G.-B., Liu S., Zhang L., et al., (2024). Building and sharing medical cohorts for research. The Innovation **5(3)**, 100623.

Cohort studies are costly and time consuming. They require not only laboratory equipment and assays but also collaboration from participants and health agencies. Due to cost constraints, they are often confined to a specific population. Nevertheless, they play a crucial role in providing fundamental insights for medical advancements, shedding light on the origins of diseases, and acting in socioeconomic influence in policy making.

A "BABY BOOM" OF COHORTS

We are currently witnessing a significant increase in the establishment of large-scale medical cohorts, covering a wide range of areas from phenomics to comprehensive multi-omics (Figure 1). Many of these cohort studies are focused on specific diseases, such as stroke[1] and occupational hazards,[2] as well as large-scale prospective studies.[3] One important development is the quick emergence of a new type of genomic data as a result of the widespread use of noninvasive prenatal testing (NIPT), which involves sequencing cell-free DNA from maternal plasma. These NIPT-derived genotypes (NIPTgs) are analogous to sparse whole-genome sequencing data, and because they accumulate so quickly, a single cohort of NIPTgs can be sampled at a far lesser cost than chipped or sequenced cohorts. The integration of electronic health records with local health information infrastructure has the potential to generate new cohorts related to NIPTgs.

The improved sequencing technique, IT technology, and relatively homogeneous population structure will make future medical cohorts (or biobanks) in China superior to traditional cohorts such as the Framingham cohort. Data will be sourced not only from research institutes but also from hospitals and smart device companies. An analogy to conceptualize emerging biobanks is as a banking system that facilitates various transactions within and between biobanks. Data exchange and interactions have been routine in medical research, but when human subjects, particularly DNA data, are involved, accomplishing data exchange becomes a major challenge. This poses a liquidity risk for the biobank system, which needs to establish trustworthiness.

The comprehensive delineation of biobank research is difficult. Our focus is directed toward genetic and genomic aspects, which are governed by strict regulations for data exchange and sharing. We investigated potential scenarios for data exchange and assessed how new technologies might address these challenges. The typical applications of cohorts, such as genome-wide association studies (GWASs) and polygenic score (PGS), depend on the sample size of the discovery dataset. Considering the genetic homogeneity of the Chinese population, merging cohorts can enhance statistical power.

ROUTINE 1: GENOME-WIDE ASSOCIATION META-ANALYSIS

Genome-wide association meta-analysis (GWAMA) has been a safe method to share summary statistics and, consequently, has facilitated statistical discoveries for genes underlying complex traits and diseases. The most recent GWAMA has involved over 5 million GWAS samples. However, one limitation of GWAMA is that the summary statistics are predefined and adjusted, potentially limiting their parameterization space. Efforts are now being made to enhance this powerful tool.

ROUTINE 2: GENOTYPE IMPUTATION

In GWAS meta-analysis, genotype imputation is imperative for maximizing statistical power. It has been a routine for genetic studies, which saturates genotypes via an *in silico* solution. High-quality reference panels and better ethnicity matching are required. Recently, the China National GeneBank Database has been certified as a Trustworthy Data Repository by the CoreTrustSeal Standards and has established such an imputation service (https://db.cngb.org/imputation/). Its reference panel is from 10,000 stroke individuals from the China Kadoorie Biobank. In addition, a meta-imputation algorithm was developed that allows imputation results generated using different reference panels to be combined into a consensus imputed dataset, which addressed the loss of power caused by privacy restrictions to some degree.[4] As imputation inevitably involves at least two datasets, to date, homomorphic encryption has been implemented to facilitate secure outsourcing of genotype imputation.

ROUTINE 3: POLYGENIC GENETIC SCORE

As single variants have small effects in determining the variation of complex traits, PGS has been extensively used to predict individual-level liability to a trait or disease. The statistical power of PGS is based on the sample size. Often, the aggregated sample size is from either single bulk data for common complex quantitative traits (such as the UK Biobank of approximately 500,000 samples) or multiple datasets for a complex disease (such as GWAMA by pooling together all available cohorts). However, our practically useful sample size is approximately 10,000 currently, so it may take STROMICS and ChinaMAP some time to collect as many samples as UK Biobank. To fully unleash the power of PGS, parameterization of a training model is often required, which in turn requires complete access to the original data or innovative computational strategies.

ROUTINE 4: LARGE LANGUAGE MODEL FOR MEDICINE

The ChatGPT has achieved remarkable success in handling diverse tasks and has demonstrated competency in the US Medical Licensing Examinations. Given the pressing need for high-quality healthcare, there is a growing interest in the development of an advanced and tailored medical GPT-like system based on natural language processing. Such a system could significantly impact clinical practice, improve efficiency, and enhance the effectiveness of clinical and educational work within the healthcare system. While there have been advancements in "few shots" or "zero shots" in large language models (LLMs), the acquisition of high-quality and high-volume medical records, including image data for a multimodal clinical LLM, remains essential. This is particularly relevant for the potential introduction of the LLM for medical applications in China, where the use of ICD-10 and ICD-11 is prevalent and anticipated to align with ICD-11 in the near future. These medical records are often centralized at national or provincial health commissions, and efforts are underway to establish or enhance the infrastructure for data sharing between the central hub and hospitals. It is imperative to recognize that LLMs play a pivotal role in current medical artificial intelligence systems, and the governance structures need to address not only the permissibility of such tools but also the potential risks, including adversarial attacks, especially in the context of chatbot-style models that could potentially harm patients.

ROUTINE 5: SEARCHING RELATIVES ACROSS COHORTS

When cohorts expand, there is an inherent interest in exploring common individuals or relatives across these cohorts for medical purposes, such as investigating kinship between donors and recipients. While there are different methods available to confirm kinship, establishing a sharing mechanism across

Figure 1. Metaphor of data sharing in medical research New medical cohorts or biobanks have mushroomed in an unprecedented way, but they are so difficult to access and share for many researchers. Are we really lacking in data or lacking the willingness for data sharing?

inter-cohort settings can be challenging. It is noteworthy that an algorithm for determining kinship using inter-cohort genomic data has been successfully developed in China recently.[5]

WHAT IMPEDES DATA SHARING

In the United States, research sponsored by the National Institutes of Health is required to release the relevant datasets or deposit them into dbGaP. In China, medical cohort funding often comes from various sources, and enforcing data sharing by one agency may not lead to widespread adoption by others. It is important to highlight the successful data sharing of the Chinese Glioma Genome Atlas, which releases medical data of various tiers (http://www.cgga.org.cn). However, many other cohort studies do not regularly practice this. Most cohorts are still in the early stages, and researchers are hesitant to share their data. As academic requirements often prioritize monopolistic authorship, data sharing consequently seems harmful if proper compartmentalization is not practiced. However, if we consider the cost for a single cohort, such as one of the aforementioned cohorts, approximately 38 million RMB is spent for sequencing, and the total expense is even magnified, say, 60 million RMB. A budget of 60 million RMB is analogous to a funding competition arena that hosts 500 applicants (0.6 million RMB and approximately 20% success rate) or 1,000 junior applicants (0.3 million RMB and approximately 20% success rate). Would it sound reasonable that a such cohort should be consequently accessed by a similar number of researchers?

A commonly raised truism is the privacy risk associated with data sharing, and there is a balance between research activities and data security. While it may seem like the cohort owner is primarily responsible for data security, it presents challenges yet also offers opportunities for the development of new technologies. Access control and authorized users are the most frequently utilized strategies, and they require fundamental infrastructure development, often guided by various national institutes. However, research activities often go beyond these established routines. Therefore, more adaptable and efficient tools are necessary. Building and sharing, which may currently seem like obstacles, are the inevitable paths for medical cohorts.

REFERENCES

1. Cheng, S., Xu, Z., Bian, S., et al. (2023). The STROMICS genome study: deep whole- genome sequencing and analysis of 10K Chinese patients with ischemic stroke reveal complex genetic and phenotypic interplay. Cell Discov. ***9***: 75. https://doi.org/10.1038/s41421-023-00582-8.
2. Du, Z., Ma, L., Qu, H., et al. (2019). Whole genome analyses of Chinese population and De Novo assembly of a northern Han genome. Genom. Proteom. Bioinform. ***17***: 229–247. https://doi.org/10.1016/j.gpb.2019.07.002.
3. Chen, Z., Chen, J., Collins, R., et al. (2011). China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. Int. J. Epidemiol. ***40***: 1652–1666. https://doi.org/10.1093/ije/dyr120.
4. Yu, K., Das, S., LeFaive, J., et al. (2022). Meta-imputation: An efficient method to combine genotype data after imputation with multiple reference panels. Am. J. Hum. Genet. ***109***: 1007–1015. https://doi.org/10.1016/j.ajhg.2022.04.002.
5. Zhang, Q.X., Liu, T., Guo, X., et al. (2024). Searching across-cohort relatives in 54, 092 GWAS samples via encrypted genotype regression. PLoS Genet. ***20***: e1011037. https://doi.org/10.1371/journal.pgen.1011037.

DECLARATION OF INTERESTS

The authors declare no competing interests.