Twin Research and Human Genetics (2019), **22**, 482–485
doi:10.1017/thg.2019.85

CAMBRIDGE UNIVERSITY PRESS

Article

The Chinese National Twin Registry: A Unique Data Source for Systems Epidemiology of Complex Disease

Tao Huang[1], Wenjing Gao[1], Jun Lv[1], Canqing Yu[1], Tao Wu[1], Shengfeng Wang[1], Chunxiao Liao[1], Lu Meng[1], Dongmeng Wang[1], Zhaonian Wang[1], Zengchang Pang[2], Min Yu[3], Hua Wang[4], Xianping Wu[5], Zhong Dong[6], Fan Wu[7], Guohong Jiang[8], Xiaojie Wang[9], Yu Liu[10], Jian Deng[11], Lin Lu[12], Weihua Cao[1,*] and Liming Li[1,*]

[1]Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing, China, [2]Qingdao Center for Disease Control and Prevention, Qingdao, China, [3]Zhejiang Center for Disease Control and Prevention, Hangzhou, China, [4]Jiangsu Center for Disease Control and Prevention, Nanjing, China, [5]Sichuan Center for Disease Control and Prevention, Chengdu, China, [6]Beijing Center for Disease Control and Prevention, Beijing, China, [7]Shanghai Center for Disease Control and Prevention, Shanghai, China, [8]Tianjin Center for Disease Control and Prevention, Tianjin, China, [9]Qinghai Center for Disease Control and Prevention, Xining, China, [10]Heilongjiang Agricultural Center for Disease Control and Prevention, Harbin, China, [11]Handan Center for Disease Control and Prevention, Handan, China and [12]Yunnan Center for Disease Control and Prevention, Kunming, China

Abstract

The Chinese National Twin Registry (CNTR), initiated in 2001, has now become the largest twin registry in Asia. From 2015 to 2018, the CNTR continued to receive Chinese government funding and had recruited 61,566 twin-pairs by 2019 to study twins discordant for specific exposures such as environmental factors, and twins discordant for disease outcomes or measures of morbidity. Omic data, including genetics, genomics, metabolomics, and proteomics, and gut microbiome will be tested. The integration of omics and digital technologies in public health will advance our understanding of precision public health. This review introduces the updates of the CNTR, including study design, sample size, biobank, zygosity assessment, advances in research and future systems epidemiologic research.

Keywords: Omic data; precision public health; systems epidemiology; twin discordance; twins

(Received 27 May 2019; accepted 14 June 2019; First Published online 11 November 2019)

A Short History

The Chinese National Twin Registry (CNTR), established in 2001, is the first national twin registry in China (Yang et al., 2002), led by the School of Public Health, Peking University, which collaborates with the Qingdao Center for Disease Control and Prevention (CDC), Dezhou CDC, Zhejiang CDC, Jiangsu CDC, Sichuan CDC, Beijing CDC, Shanghai CDC, Tianjin CDC, Qinghai CDC, Heilongjiang agricultural area CDC, Handan CDC, Yunnan CDC and Harbin Medical University. Professor Liming Li from Peking University is the principle investigator.

The first financial support was from an independent American foundation, the Rockefeller-endowed China Medical Board, from 2001 to 2005. At the very beginning, the CNTR selected Qingdao, Beijing, Shanghai and Lishui as the first four cities to recruit twins. In 2006, an article introducing the CNTR was published in a special issue of twin registries (Li et al., 2006). In 2010, the CNTR obtained a grant from the Chinese government (the Special Fund for Health Scientific Research in Public Welfare, 201002007), which expanded the twin registry from four areas to nine provinces or cities (p/c; Li et al., 2013). From 2015 to 2018, the CNTR continued to receive Chinese government funding (201502006) and had recruited 61,566 twin-pairs (including multiple sets) in 11 p/c by February 2019. The demographic information on twins is presented in Table 1. The CNTR has now become the largest twin registry in Asia. The CNTR recently celebrated its 20th anniversary on January 19, 2019.

Study Design

Twin Recruitment

The CNTR is a voluntary registry. The twins are mainly identified and recruited through the local CDC, which shares data with the residence registry in local public security bureaus and communities, or through public media. The CNTR collects twin data through face-to-face interviews with investigators from the CDC, which covers all levels of province, city and county in China. The CNTR also recruits twins through the Chinese 'Hukou' system (an ID schema), which is administered by the public security bureau. Data from the CDCs and the public security bureau are used to identify twins, which are verified by health workers. In addition, advertisements in print or online media are also used. The study protocol for the Special Cohort Study on Environmental Epidemiology in China was reviewed and approved by the Ethics Committee for Human Subject Studies of the Peking University Health Science Center in 2014 (ID: IRB000 01052-14021).

Author for correspondence: Liming Li, Email: lmlee@vip.163.com
*These authors contributed equally to this work.

Cite this article: Huang T, Gao W, Lv J, Yu C, Wu T, Wang S, Liao C, Meng L, Wang D, Wang Z, Pang Z, Yu M, Wang H, Wu X, Dong Z, Wu F, Jiang G, Wang X, Liu Y, Deng J, Lu L, Cao W, and Li L. (2019) The Chinese National Twin Registry: A Unique Data Source for Systems Epidemiology of Complex Disease. *Twin Research and Human Genetics* **22**: 482–485, https://doi.org/10.1017/thg.2019.85

Table 1. Distribution of twin-pairs by zygosity, gender and age at recruitment

	Age at recruitment										
	0–	5–	10–	15–	20–	30–	40–	50–	60–	Missing	Total
MZ											
MZM	2336	1764	1363	1423	3662	2582	2368	1196	705	1	17,400
MZF	2283	1722	1272	1502	3766	1723	1220	504	228	2	14,222
DZ											
DZM	1496	912	703	648	1856	1458	1073	528	272	2	8948
DZF	1260	762	657	523	1144	945	471	214	78	1	6055
DZO	2763	1677	1201	1168	3055	1796	1122	516	198	4	13,500
Triplet/quadruplet sets											
MZ	21	26	17	6	10	1	2	0	0	0	83
Same-sex DZ	11	19	8	3	9	6	0	1	0	0	57
DZO	5	1	0	5	14	3	3	0	0	0	31
Others	312	243	161	119	143	122	81	45	41	3	1270
Total	10,487	7126	5382	5397	13,659	8636	6340	3004	1522	13	61,566

Note: MZ = monozygotic twins, DZ = dizygotic twins, M = male, F = female, O = opposite sex. Others include twin-pairs/multiple sets with gender or PPQ information missing, or twins who answered PPQ question with 'I have no idea'.

Zygosity Assessment

We used the Peas in the Pod Questionnaire (PPQ) for twin zygosity assessment when genotyping information is not available. A total of 1008 twin-pairs recruited in 2001 was assessed by genetic test. In China, according to our own comparison studies, the accuracy of twin diagnosis can reach 86–90% (Li et al., 2013). There were no statistically significant sex and area differences in the validity of the questionnaire and physical features comparison-based classification. Therefore, questionnaire-based zygosity assessment in this Chinese adult twin sample can still be regarded as a valid and valuable classification method. Physical features comparisons, however, can only provide limited information for zygosity determination (Gao et al., 2006).

As for twins for whom blood samples are available, we use several methods, including blood group typing, analysis of four or nine short tandem repeat genetic markers, single-nucleotide polymorphisms (SNPs) using the Illumina HumanOmniZhongHua-8 BeadChip, SNP information from the Illumina Infinium Human Methylation 450 BeadChip and MethylationEPIC '850K' BeadChip array. In our study, we recruited 192 same-sex Chinese adult twin-pairs to evaluate the validity of using the genetic marker-based method and questionnaire-based method for zygosity determination. We considered the relatedness analysis based on more than 0.6 million SNP genotyping as the gold standard for zygosity determination. The results of zygosity determination based on 65 SNPs in 450k methylation array were all consistent with genotyping. For cost considerations for twin studies with genotyping and/or 450k methylation array, there is no need to conduct other zygosity testing (Wang et al., 2015).

Twin Questionnaire

The twin data were collected through face-to-face interviews with investigators and questionnaires that were completed by interviewees. The CNTR twins were asked to complete one of the two questionnaires: a simple questionnaire for twins under 18 years and a complex questionnaire for adults ≥18 years. The first questionnaire for twins aged below 18 years collected demographic information, parents' names, birth weight, current weight and height, medical history and zygosity. Adult twins aged 18 years and over completed a more complex questionnaire that includes questions on demographic information, socioeconomic status, birth weight, birth defects, birthplace, whether reared apart, current height, weight, waist circumference, zygosity, smoking, drinking, fruit and vegetable consumption, physical activity, medical history, allergic history and family medical history. Twins are currently followed up and their contact information, height, weight and waist measurements, and the onset of new chronic diseases are updated when possible.

The CNTR Biobank

Blood Sampling

Fasting blood samples from 1008 twin-pairs, regardless of their illness, were collected during 2001–2002. Only 579 twin-pairs were followed up during 2004–2005. Basic biochemical tests and DNA extraction were performed. From 2010, we collected further blood samples from 1196 disease-discordant pairs and 577 disease-concordant pairs. The number of twin-pairs who provided a blood sample is presented in Table 2. Serum lipids, glucose, glycosylated hemoglobin, insulin, high-sensitive C-reactive protein and creatinine were tested in these serum samples.

Genotyping and Quality Control

A total of 480 twins of DNA samples were assessed for integrity, quantity and purity by electrophoresis and NanoDrop measurements. A genomewide genotyping scan among 240 pairs was carried out using Illumina HumanOmniZhongHua-8 BeadChip. High-quality genotyping was performed by laboratory specialized in Illumina SNP array genotyping following standard experimental procedures suggested by the manufacturer. A total of 894,956 SNPs was genotyped among 240 paired subjects. Genomewide heterozygosity of each individual was estimated to exclude cross-contamination between samples. This test was performed using a subset of 155,588

Table 2. The disease distribution of twin blood samples (pairs)

	MZ	DZ[1]	Zygosity unknown	Total
Disease-discordant				
Obesity	123	168	1	292
Hypertension	189	140	0	329
Diabetes	100	65	0	165
Hyperlipidemia	106	51	1	158
Coronary heart disease	53	24	1	78
Chronic bronchitis/emphysema	51	29	0	80
Cancer	30	21	0	51
Stroke	29	14	0	43
Disease-concordant				
Obesity	88	40	0	128
Hypertension	212	67	1	280
Diabetes	73	22	0	95
Hyperlipidemia	20	16	1	37
Coronary heart disease	14	3	0	17
Chronic bronchitis/emphysema	12	3	0	15
Cancer	0	1	0	1
Stroke	3	1	0	4
Twins reared apart				
	88	52	1	141

Note: Obesity: BMI $\geq$ 28 kg/m^2 based on height and weight measured. Other diseases are based on self-reported diagnosis by county-level hospital or above, or taking therapeutic drugs for hypertension/diabetes/hyperlipidemia/cancer. Twins reared apart: being separated at least 1 year before 11 years old.
[1]Includes both same-sex and opposite-sex twin-pairs.

SNPs pruned for linkage disequilibrium ($r^2 < .3$). Two samples demonstrated signs of contamination (mean observed heterozygosity = .3404, standard deviation = .0052), and those two samples and their twin siblings were excluded. Finally, 180 pairs of twins with 695,406 SNPs remained after the quality control filters.

Genomewide Methylation Profiling

Overall, 118 monozygotic (MZ) and 97 dizygotic (DZ) twin-pairs were tested using Illumina 450k methylation array, and 87 MZ and 51 DZ twin-pairs using Illumina 850k methylation array in the CNTR. DNA methylation level was displayed as beta-values ranging from 0 to 1. Beta-value was defined as the algorism M/(M + U + 100), where M and U represent the methylated and unmethylated signal intensities, respectively. For the methylation quality control, we used the 65 SNPs to determine the genotype of the sample and compare it with genotype calls based on Zhonghua8 Beadchip data, to identify if there were some mix-ups during the methylation experiment. Sample-level and probe-level quality control showed that all samples passed the Illumina quality control. About 1% of sites with a detection p value greater than .01 were removed (zero sample). Sites having 1% of samples with a detection p value greater than .01 (4019 sites) or sites with bead-counts <3 in 5% of samples (1499 sites) were removed. In addition, sites with SNPs or with a minor allele frequency of at least 5% were excluded, because probe binding might be affected by SNPs in the binding area. Finally, a total of 817,471 probes that passed the quality control were included.

Distribution of Discordant or Concordant Twin-Pairs

The co-twin control study design involves twins discordant for specific exposures such as environmental factors, and twins discordant for disease outcomes or measures of morbidity.

Disease-discordant twins, particularly MZ twins, are excellent subjects for matched case-control studies in which confounding effects of age, sex, genetic background, intrauterine and early environments are perfectly controlled. Therefore, analyzed differences in outcome against differences in exposure, within- and between-pair models, and conditional logistic regression can be used for potential causal inference. Thus, both disease-discordant and disease-concordant twin-pairs are invaluable resources.

At the CNTR, our research mainly focuses on cardiometabolic diseases such as diabetes, obesity, hypertension and genetic diseases among children or adolescent twins; and cardiometabolic diseases, chronic bronchitis/emphysema and cancers in adult twins. At the CNTR, obesity is self-reported after diagnosis by county-level hospital or above for twins younger than 18 years; for twins $\geq$18 years, obesity is BMI $\geq$ 28 kg/m^2 based on self-reported height and weight. All other diseases are self-reported using a questionnaire. The top three discordant diseases are obesity, hypertension and diabetes. The distribution of disease or lifestyle-discordant and -concordant twin-pairs is listed in Table 3.

Advances of the Twin Cohort

Both genetic and environmental factors contribute to cardiometabolic health. The CNTR comprehensively examined the genetic and environmental effects on variances in weight, height and body mass index (BMI) under 18 years. We found that heritability for weight, height and BMI was low at 0–2 years old (less than 20% for both sexes) but increased over time. Therefore, genetics appear to play an increasingly important role in explaining the variation in weight, height and BMI from early childhood to late adolescence, particularly in boys. Common environmental factors exert their strongest and most independent influence specifically in the pre-adolescent period and more significantly in girls (Liu et al., 2015). Our findings emphasize the need to target family and social environmental interventions in early childhood years, especially for females (Liao et al., 2018). We further quantified and compared the associations of various body composition measurements with serum metabolites and to what degree genetic or environmental factors affect obesity-metabolite relation. Adiposity showed significant associations with serum metabolite concentrations. Of these phenotypic correlations, 64–81% was attributed to genetic factors, whereas 19–36% was attributed to unique environmental factors. To a large degree, shared genetic factors contributed to these associations, with the remainder explained by twin-specific environmental factors (Liao et al., 2015). Interestingly, a structural equation model adjusting for age and sex found vigorous exercise significantly moderated the additive genetic effects and shared environmental effects on BMI. The genetic contributions to BMI were significantly lower for people who adopted a physically active lifestyle ($h^2 = 40\%$) than those who were relative sedentary ($h^2 = 59\%$). The observed gene–physical activity interaction was more pronounced in men than in women. Therefore, our finding suggests that adopting a physically active lifestyle may help to reduce the genetic influence on BMI among the Chinese population (Wang et al., 2016).

Table 3. Disease- and lifestyle-discordant and disease- and lifestyle-concordant twin-pairs at the CNTR

	Y/Y				Y/N			
	MZ	(%)[1]	DZ	(%)	MZ	(%)	DZ	(%)
Diseases								
Obesity	653	1.20	297	0.50	748	1.30	1202	2.20
Hypertension	487	0.80	210	0.40	610	1.00	584	1.00
Diabetes	175	0.30	60	0.10	256	0.40	290	0.50
Hyperlipidemia	134	0.30	64	0.20	299	0.70	266	0.70
Coronary heart disease	75	0.20	24	0.10	185	0.50	148	0.40
Chronic bronchitis/ emphysema	68	0.20	33	0.10	160	0.40	149	0.40
Cancer	45	0.10	22	0.10	149	0.40	126	0.30
Genetic disease	94	0.20	44	0.10	66	0.10	53	0.10
Stroke	38	0.10	20	0.00	109	0.30	75	0.20
Asthma	42	0.20	23	0.10	50	0.20	94	0.40
Diet and lifestyle								
Smoking[2]	3267	10.10	1430	4.40	1,751	5.40	3432	10.60
Alcohol drinking	2712	8.40	1367	4.20	1,471	4.60	2671	8.30
Fruit and vegetable consumption[3]	5249	18.80	3913	14.00	623	2.20	711	2.50
Physical activity[4]	5740	19.40	4278	14.40	1906	6.40	2086	7.00

Note: Y = disease present; N = disease absent. Obesity: For twins <18 years, obesity is self-reported after diagnosis by county-level hospital or above; for twins ≥18 years, obesity is BMI ≥ 28 kg/m^2 based on self-reported height and weight.
[1]Percentage of discordant twin-pairs among whole population.
[2]Y = current smoker/drinker; N = never smoker/drinker or ex-smoker/drinker. A current smoker is defined as anyone who, self-reportedly, smokes one or more cigarettes (or cigars, pipes or any other smoked tobacco products) daily in the past year. The definition of a current drinker is anyone who self-reportedly consumes >50 g of liquor with 52% alcohol by volume daily in the past year.
[3]Y = those who eat at least three servings of vegetables and two servings of fruit per day.
[4]Y = those who do moderate or vigorous physical activity for at least 30 min at a time on at least 5 days per week. Moderate activities refer to activities that take moderate physical effort and make breathing somewhat harder than normal. Vigorous physical activities refer to activities that require hard physical effort and make breathing much harder than normal.

Future Plans and Systems Epidemiology

Remarkable advances in omics technologies, including genomics, metabolomics and proteomics, and the study of the human microbiome have provided many new opportunities for epidemiologic research. The integration of such omic technologies into epidemiology by adopting a 'systems epidemiology' approach can help to understand the etiology of chronic diseases, discover novel biomarkers and identify high-risk populations to target for precision intervention.

In the next step, more fasting blood samples will be collected in the disease/exposure-discordant twins, who will undergo a detailed physical examination. Matched case-control and cohort studies will be conducted in these discordant twins. Omic data, including genetics, genomics, metabolomics and proteomics, and gut microbiome will be tested. The integration of omics and digital technologies in public health will advance our understanding of precision public health. Therefore, a research agenda that incorporates a multidisciplinary approach applied across the life cycle, which leverages new technologies and which addresses new challenges should be most effective for the prevention of chronic diseases.

Aims of the Twin Cohort

The CNTR aims to investigate the genetic and environmental contributions to complex diseases, with particular emphasis on cardiovascular diseases. During the past 20 years of growth, however, the CNTR has not only provided important insights into cardiovascular diseases but also served as a valuable resource for a broad range of study areas. The CNTR has collected data on health, lifestyle and behavior, as well as fasting blood samples. Longitudinal follow-ups and surveillance of major chronic diseases are also planned. Therefore, our research will extend to examine the causal relationship between environmental risk factors and chronic diseases using the disease-discordant twins and exposure-discordant twins.

With the development of diverse high-throughput technologies, the rapidly evolving field of omic data offers the potential to study health and disease in breadth and depth at the human population level. Therefore, we also plan to use systems epidemiology as a novel approach to study the complexities of human pathophysiology or identify new trans-omic biomarkers by integrating genetics, gut microbiota, epigenetics, genomics, metabolomics and population phenotype data.

Acknowledgments. The CNTR is supported by the special fund for health scientific research in public welfare, China (201002007, 201502006), Key Project of Chinese Ministry of Education (310006), National Natural Science Foundation of China (81573223, 81473041, 81202264, 81711530051) and China Medical Board (01-746). We gratefully acknowledge support from the Centers of Disease Control and Prevention in Qingdao, Dezhou, Zhejiang, Jiangsu, Sichuan, Beijing, Shanghai, Tianjin, Qinghai, Heilongjiang agricultural area, Handan, and Yunnan, and School of Public Health, Harbin Medical University.

Reference

Gao, W., Li, L., Cao, W., Zhan, S., Lv, J., Qin, Y., … Hu, Y. (2006). Determination of zygosity by questionnaire and physical features comparison in Chinese adult twins. *Twin Research and Human Genetics*, *9*, 266–271.

Li, L., Gao, W., Lv, J., Cao, W., Zhan, S., Yang, H., & Hu, Y. (2006). Current status of the Chinese National Twin Registry. *Twin Research and Human Genetics*, *9*, 747–752.

Li, L., Gao, W., Yu, C., Lv, J., Cao, W., Zhan, S., … Hu, Y. (2013). The Chinese National Twin Registry: An update. *Twin Research and Human Genetics*, *16*, 86–90.

Liao, C. X., Gao, W. J., Cao, W. H., Lv, J., Yu, C. Q., Wang, S. F., … Li, L. M. (2015). Associations of body composition measurements with serum lipid, glucose and insulin profile: A Chinese Twin Study. *PLoS One*, *10*, e0140595.

Liao, C. X., Gao, W., Cao, W., Lv, J., Yu, C., Wang, S., … Li, L. (2018). Association of educational level and marital status with obesity: A study of Chinese twins. *Twin Research and Human Genetics*, *21*, 126–135.

Liu, Q., Yu, C., Gao, W., Cao, W., Lyu, J., Wang, S., … Li, L. (2015). Genetic and environmental effects on weight, height, and BMI under 18 years in a Chinese population-based twin sample. *Twin Research and Human Genetics*, *18*, 571–580.

Wang, B., Gao, W., Lv, J., Yu, C., Wang, S., Pang, Z., … Li, L. (2016). Physical activity attenuates genetic effects on BMI: Results from a study of Chinese adult twins. *Obesity (Silver Spring)*, *24*, 750–756.

Wang, B., Gao, W., Yu, C., Cao, W., Lv, J., Wang, S., … Li, L. (2015). Determination of zygosity in adult Chinese twins using the 450K methylation array versus questionnaire data. *PLoS One*, *10*, e0123992.

Yang, H., Li, X., Cao, W., Lu, J., Wang, T., Zhan, S., … Li, L. (2002). Chinese National Twin Registry as a resource for genetic epidemiologic studies of common and complex diseases in China. *Twin Research*, *5*, 347–351.