Published by Oxford University Press on behalf of the International Epidemiological Association
International Journal of Epidemiology 2011;40:1652–1666
doi:10.1093/ije/dyr120

China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up

Zhengming Chen,[1]* Junshi Chen,[2] Rory Collins,[1] Yu Guo,[2,3] Richard Peto,[1] Fan Wu,[2,4] and Liming Li[2,3,5,]* on behalf of the China Kadoorie Biobank (CKB) collaborative group[†]

[1]Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU), Nuffield Department of Clinical Medicine, University of Oxford, Oxford, UK, [2]Chinese Centre for Disease Control and Prevention, Chang Ping District, Beijing, China, [3]Chinese Academy of Medical Sciences, Dong Cheng District, Beijing, China, [4]Shanghai Municipality Centre for Disease Control and Prevention, Shanghai, China and [5]School of Public Health, Beijing University, Beijing, China

*Corresponding author. CTSU, Nuffield Department of Clinical Medicine, University of Oxford, Old Road Campus, Oxford OX3 7LF, UK. E-mail: zhengming.chen@ctsu.ox.ac.uk, or Chinese Academy of Medical Sciences, Dong Cheng District, Beijing 100730, China. E-mail: lmlee@vip.163.com

†The members of the steering committee and collaborative group of the China Kadoorie Biobank are listed in the Acknowledgements section

Accepted 7 July 2011

Background Large blood-based prospective studies can provide reliable assessment of the complex interplay of lifestyle, environmental and genetic factors as determinants of chronic disease.

Methods The baseline survey of the China Kadoorie Biobank took place during 2004–08 in 10 geographically defined regions, with collection of questionnaire data, physical measurements and blood samples. Subsequently, a re-survey of 25 000 randomly selected participants was done (80% responded) using the same methods as in the baseline. All participants are being followed for cause-specific mortality and morbidity, and for any hospital admission through linkages with registries and health insurance (HI) databases.

Results Overall, 512 891 adults aged 30–79 years were recruited, including 41% men, 56% from rural areas and mean age was 52 years. The prevalence of ever-regular smoking was 74% in men and 3% in women. The mean blood pressure was 132/79 mmHg in men and 130/77 mmHg in women. The mean body mass index (BMI) was 23.4 kg/m^2 in men and 23.8 kg/m^2 in women, with only 4% being obese (>30 kg/m^2), and 3.2% being diabetic. Blood collection was successful in 99.98% and the mean delay from sample collection to processing was 10.6 h. For each of the main baseline variables, there is good reproducibility but large heterogeneity by age, sex and study area. By 1 January 2011, over 10 000 deaths had been recorded, with 91% of surviving participants already linked to HI databases.

Conclusion This established large biobank will be a rich and powerful resource for investigating genetic and non-genetic causes of many common chronic diseases in the Chinese population.

Keywords Biobank, cohort studies, chronic diseases, aetiology, China

Introduction

Chronic non-communicable disease such as ischaemic heart disease (IHD), stroke and cancer are now substantial causes of death and disability in low- and middle-income countries, such as China.[1,2] At current death rates, 17 of the 18 million children born each year in China will survive until age 35 years, but ~7 million will then die in middle age (35–69 years), mainly due to various chronic diseases.[3–5] Population mortality rates from many common chronic diseases such as IHD and lung cancer have been declining in most Western countries, but they are increasing in China as a result of adverse changes in lifestyle, diet and tobacco use.[5–7] For each major chronic disease, there is still large unexplained variation in the age-specific rates between different populations and, within China, between different regions,[8–10] suggesting that some important non-genetic causes remain to be discovered. In addition, genetic factors and gene–environment interactions are likely to play important roles in disease causation.[11–13] Although case–control studies may suffice for the study of purely genetic factors, large blood-based prospective cohort studies are essential for the unbiased assessment of the relevance of both genetic and environmental factors, and their interactions, as determinants for common chronic diseases.

Several important causes of various chronic diseases are already known,[14–17] but this knowledge is mainly based on studies in the West and does not generally suffice to explain much of the large geographic differences in disease rates around the world and between different regions of China.[8–10,18] In part, this may reflect the unreliability of present estimates of the age-specific importance of prolonged exposure to known risk factors in different populations, particularly when they are acting in combination with each other.[15–17] Although there have already been several prospective studies of major chronic diseases in China,[19–25] each has had its limitations, including insufficient numbers,[19,22,24] lack of blood samples,[20,21] involving just one city or occupational cohort[22–24] and limited data collection on risk exposures and outcome measures.[19–21] Consequently, the aetiology of many common chronic diseases in China is still poorly understood, and there is still substantial uncertainty about the present and future relevance to population mortality of many common risk factors, such as smoking.[20,26] In 2004, we launched a large blood-based prospective study, the China Kadoorie Biobank [CKB, known previously as the Kadoorie Study of Chronic Disease in China (KSCDC)], with the goal of recruiting and assessing 0.5 million people and then following their health for at least 2 decades.[27] We report here the detailed survey methods, the main baseline characteristics of the participants and status of follow-up.

Materials and Methods

Baseline survey

The study took place in 10 geographically defined regions (5 urban and 5 rural) of China, chosen according to local disease patterns, exposure to certain risk factors, population stability, quality of death and disease registries, local commitment and capacity. For the baseline survey, a Regional Coordinating Centre (RCC) and survey team, consisting of about 15 full-time staff with medical qualifications and fieldwork experience, were established in each of the 10 study areas. Potentially eligible participants in each of 100–150 administrative units (rural villages or urban residential committees) selected for the study within each region were identified through official residential records, and invitation letters (with study information leaflets) were delivered door-to-door by local community leaders or health workers, following extensive publicity campaigns. As a pre-requisite for participating, all participants were asked to bring their unique national identity (ID) cards to the assessment centre set up in the local community. To encourage participation, any individuals attending the baseline survey who were slightly outside the target age range (35–74 years) were not turned away.

After registration and giving informed written consent (which allows access to their medical records and long-term storage of blood for anonymized and non-specified medical research purposes), each participant moved through various stations in the assessment centre. The whole clinic visit typically took ~60–75 min to complete. The target daily recruitment rate was 70–80 participants per region (i.e. overall daily recruitment of 700–800 participants).

The interviewer-administered electronic questionnaire consisted of 10 major sections related to general demographic and socio-economic status, dietary and other lifestyle habits (e.g. smoking, alcohol and tea drinking), exposure to passive smoking and domestic indoor air pollution, medical history and current medication, physical activity, sleeping and mental status (using CIDI-SF[28]) and reproductive history (for women). The physical measurements included height, weight, hip and waist circumference, bio-impedance, lung function, carbon monoxide (CO), blood pressure and pulse rate, using standard instruments and protocols and with regular calibrations.[27] In addition, blood spot tests were also done on random blood glucose and hepatitis B surface antigen (but not on any other blood-related markers).[27]

For each participant, a 10-ml non-fasting blood sample (with time of last meal recorded) was collected into one EDTA vacutainer (BD Hemogard™, USA). The samples were then kept in a portable, insulated cool box with ice packs (to maintain their temperature at 0–4°C) for up to a few hours before being taken to the local study laboratory for immediate processing. After centrifuging and aliquoting, the four

cryovials (including one DNA-containing buffy coat) from each blood sample were stored in a −40°C freezer for 3–4 months, before being couriered on dry ice to the central blood repository in Beijing for storage at −80°C. Every 6 months, two frozen aliquots of plasma sample from each participant were couriered on dry ice from Beijing to Oxford for long-term storage in liquid nitrogen tanks.

Within a few weeks of the initial baseline survey in a particular community (e.g. village), a quality control (QC) survey was done, involving ~3% of the participants randomly selected from that community with repeat questionnaire and measures on selected items. During the course of the survey, regular central monitoring was also undertaken to assess the recruitment rate, the distribution of certain key variables, the time delay with blood processing and consistency of the data collected, both overall and by individual staff. On-site monitoring visits were also undertaken every 6 months by provincial Centres for Disease Control and Prevention (CDC) staff and annually by Oxford/Beijing staff.

Overall, a total of 354 local staff were involved in the baseline survey and the total number who completed 95% of all the data collection for the whole study was 21 for initial registration and consenting, 20 for anthropometric measurements, 24 for blood pressure measurements and blood collection, 23 for lung function and CO and 113 for the main questionnaire.

Prior to starting the project, central ethics approvals were obtained from Oxford University, and the China National CDC. In addition, approvals were also obtained from institutional research boards at the local CDCs in the 10 regions.

Re-survey

Following completion of the baseline survey in July 2008, a re-survey was undertaken in 10 study regions during August and October 2008. It involved ~5% of randomly chosen surviving participants and used administrative unit (i.e. rural village or urban residential committee) as the basic sampling unit. Apart from a few additional questions (e.g. recent hospitalization), the data collection and survey procedures were much the same as in the baseline survey. The data from this first re-survey (as well as the planned subsequent periodic re-surveys) will allow, after controlling for 'regression dilution' bias,[29] unbiased prospective assessment of associations between long-term 'usual' levels of particular risk exposures and disease in the whole population. For the subsequent re-surveys, new and more detailed data collection to the study (e.g. better measures of physical activity or dietary patterns) will be considered, so that information collected at the baseline survey can be calibrated and enhanced.[30]

Long-term follow-up

All 10 study regions are part of China's Disease Surveillance Points (DSP) system, which provides mortality statistics for the entire country.[3] The vital status of study participants is being monitored regularly through official residential records and death certificates reported to the Regional CDC, where each study RCC office is based. Any deaths occurring among participants are coded (using the 10th International Classification of Diseases, ICD-10) by trained staff 'blinded' to baseline information. Causes of death from official death certificates are being supplemented, if necessary, by review of medical records (which are usually available). To help minimize the under-reporting of deaths and to identify participants who have moved permanently out of the study areas, separate active confirmation of vital status is also being carried out annually by reviewing of residential records and/or by visits to local communities. For any additional deaths not identified through routine procedures, the causes will be sought by reviewing hospital records or by conducting a verbal autopsy using a validated instrument.[31]

Information on disease incidence for stroke, IHD, cancers and diabetes is also being collected through linkage with established disease registries that is currently available in 8 out of the 10 study areas. Future follow-up for incident cases of these conditions, as well as for many other types of hospital admission, will be based chiefly on electronic linkage with the new national health insurance (HI) claim databases that are now fully established in all 10 study regions. Based on the initial linkages done in study areas, personal information can be matched to these HI databases for most of the participants using the unique national ID, and ~4–5% of the participants had had at least one hospital admission recorded each year. The main phase of data collection through local HI databases is now starting in all study areas. To help improve the accuracy of diagnosis and phenotyping of reported conditions, outcome adjudication and further data collection for specific disease or diagnostic information (e.g. oestrogen receptor status for breast cancer) will also be undertaken for a range of major diseases.

Statistical analysis

For baseline variables, the mean values and prevalences were calculated separately for men and women, standardized by 5-year age group and areas. Where necessary, data were also analysed separately for urban and rural areas. For agreement between baseline and subsequent data collected at QC survey and re-survey, weighted kappa (κ) or Spearman correlation analyses were used. All analyses were conducted using SAS version 9.2.

Results

Overall, 515 681 people attended the baseline survey between June 2004 and July 2008, of whom 261 (0.05%) withdrew before completion, 2208 (0.4%) were found subsequently to have inadvertently attended the survey twice at different time points and 1 had major data errors. The estimated population response rate was ~30% (26–38% in the five rural areas and 16–50% in the five urban areas). Of the 513 211 participants with valid baseline data (i.e. completed questionnaire, physical measurements and consent form), blood collection was successful in 99.98% and the mean delay from blood sample collection (with immediate chilling) to sample separation (with immediate freezing) was 10.6 h.

Overall, of the 512 891 participants aged 30–79 years (i.e. excluding 320 outside this age range) for the present analysis, 41% of the participants were men, 56% were from rural areas and the mean age was 52 years. The number recruited from each area ranged from about 30 000 in Haikou to over 63 000 in Henan (Figure 1). There was a significant deficit in the number of participants born in 1958–61, which coincides with the great famine in China, compared with previous or subsequent years (Figure 2), and this pattern was seen in all regions. Tables 1 and 2 show the baseline characteristics of the study participants. Nearly all were married and the proportion having no spouse (mainly widowed) was more than twice as high in women as in men (11.3 vs 4.9%), reflecting higher death rates in men than in women. About half of the participants had at least 6 years of formal education, and the proportion was much higher in men than in women and in urban than in rural regions and varied significantly by area and year of birth (Figures 3 and 4). At baseline, >80% had basic health cover, and there has been further increase in the coverage rate during the subsequent years following the health-care reform. The prevalence of

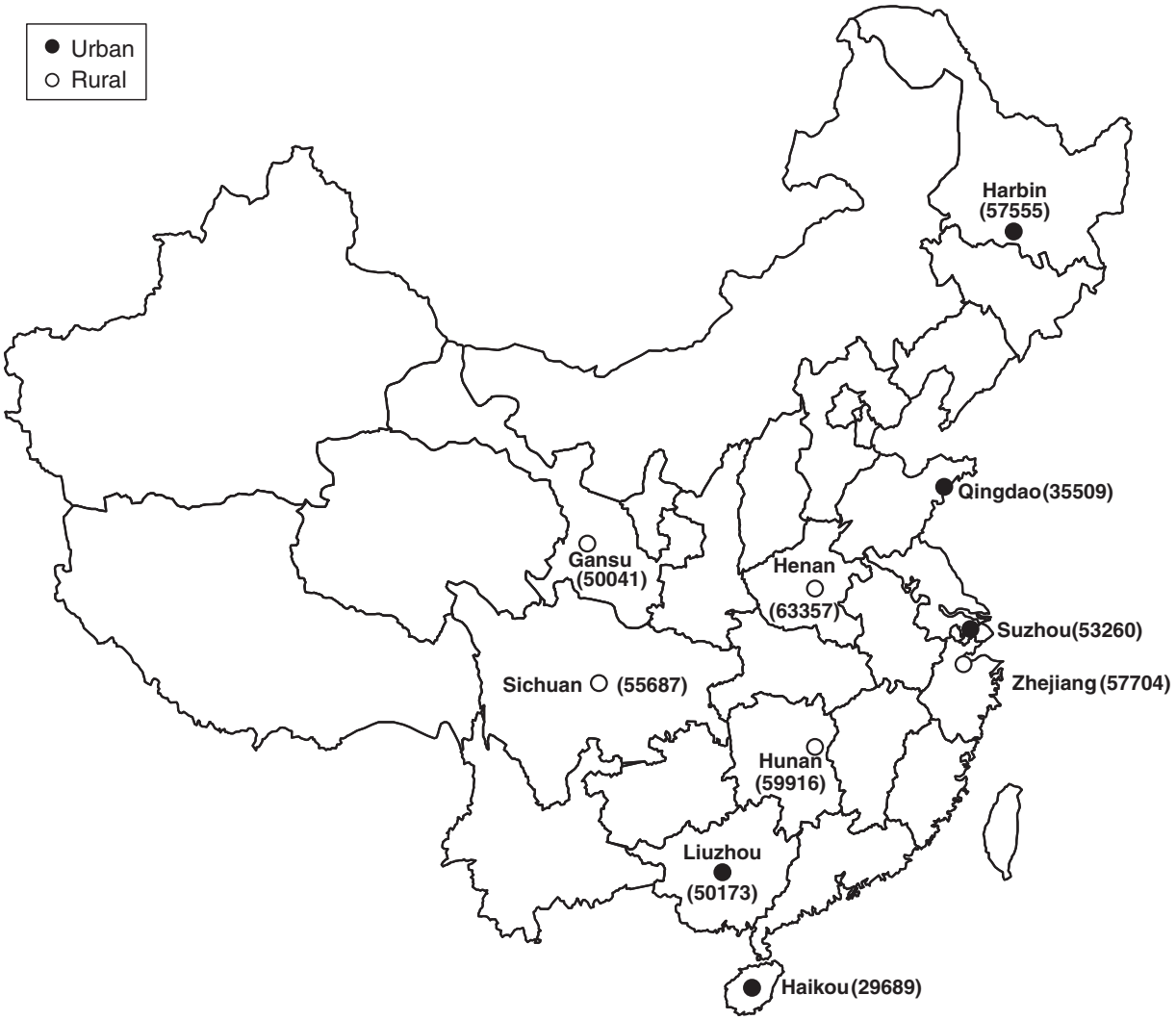

Figure 1 Locations of the 10 survey sites and number recruited. Open circles indicate rural areas and solid circles indicate urban areas. Number recruited at baseline in each area is shown in brackets

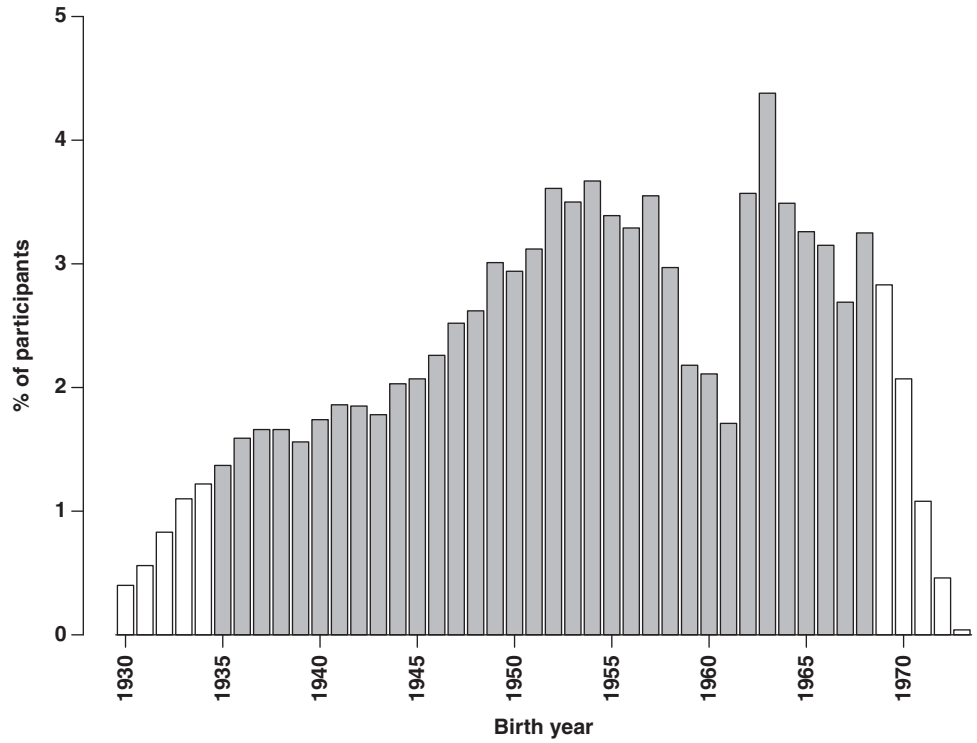

Figure 2 Proportion of participants born in each calendar year. Open bars indicate people born during these years were eligible for only part of the 4-year recruitment period at baseline survey due to age restriction. The effects of the 1959–61 famine vary by site. The small numbers of participants born before 1930 or after 1973 are shown as 1930 or 1973, respectively

ever-regular smoking was much higher in men than in women (74 vs 3%); among men, it varied little by area and age (Figures 3 and 4) but among women, there has been a major decrease since 1950 in smoking uptake rates, so the habit is extremely uncommon (<1%) in women born since 1950 (Figure 4). For regular alcohol and tea drinking (mainly green tea), large differences in prevalence were also seen between men and women and, among men, the prevalence varied greatly by area (Figure 3).

At baseline, nearly all reported daily consumption of fresh vegetables, whereas the proportions consuming meat/poultry, fish/sea food, egg, soybean products, fresh fruit on most days were 47, 9, 25, 10 and 28%, respectively (Table 1). Although only 12% reported consumption of dairy products, there was >10-fold variation across the study areas, as is the case for the consumption of spicy food (Figure 3). Overall, 17% of the participants reported taking food supplements (e.g. multi-vitamins) regularly and, in both sexes, this percentage increased steadily with higher age (Figure 4). About half reported using smoky cooking fuel (stalks, wood or coal) and the proportion was almost 10 times as high in rural as in urban areas (Figure 3).

Only about 5% of the women reported having menarche before the age of 13 years, but the proportion increased almost linearly with year of birth, with exception of an anomalous drop among women born in 1940–45, reflecting delayed menarche due to the 1959–61 famine (Figure 4). Nearly all of the women in the study had given birth, and 6.6% reported having had five or more live births, which was significantly higher in rural than in urban areas and among older women (Figure 4) and varied greatly across 10 areas even at old age (Figure 3). Only about 10% of the women reported ever using oral contraceptives, with the proportion ranging from <1 to ~40% at reproductive ages in different areas (Figure 3).

The overall estimated total daily energy expenditure (based on type, strength and duration of each type of work, transportation, exercise and household-related work[32,33] was about 26 Metabolic Equivalent of Task (MET) hours, with little difference between men and women (Table 2). Mean body mass index (BMI) was 23.4 kg/m^2 in men and 23.8 kg/m^2 in women, with only 4% being obese (>30 kg/m^2). Across the 10 study areas, the prevalence of obesity ranged from 1.6 to 11% (Figure 3). Although stroke rates are high in China, the mean blood pressure at baseline was only 132/79 mmHg in men and 130/77 mmHg in women. With the exception of hypertension, a small proportion of participants also reported having various

Table 1 Selected characteristics of study participants at baseline survey, 2004–08

Characteristics[a]	Men ($n=210\,222$), %	Women ($n=302\,669$), %	Total ($n=512\,891$), %
Age (years)			
30–39	14.1	15.9	15.2
40–49	28.2	30.9	29.8
50–59	30.3	31.0	30.7
60–69	19.7	16.7	17.9
70–79	7.8	5.5	6.4
Mean (SD)	52.8 (10.9)	51.5 (10.5)	52.0 (10.7)
Marital status			
Married with spouse	93.6	88.5	90.6
Widowed, separated, divorced	4.9	11.3	8.7
Never married	1.5	0.2	0.7
Highest education completed			
No formal school	7.5	26.3	18.6
Primary school	33.0	31.7	32.2
Middle school	33.6	24.6	28.3
High school	18.0	13.1	15.1
College or university	8.0	4.4	5.9
Socio-economic status			
Household income $\geqslant$20 000 Yuan/year	46.0	40.4	42.7
Having health cover	85.0	80.2	82.2
Having landline or mobile phone	89.5	88.4	88.9
Having flushing toilet in home	50.7	50.5	50.6
Having holiday in last 5 years	9.4	9.8	9.6
Smoking history			
Never	14.4	95.0	61.9
Occasional	11.3	1.8	5.7
Ever regular	74.4	3.2	32.4
Current regular	61.2	2.3	26.4
Regular drinking of beverages			
Alcohol (weekly)	33.3	2.0	14.8
Tea (on most days)	46.8	18.7	30.2
Regular consumption of certain foodstuffs			
Fresh vegetables	98.3	98.3	98.3
Meat/poultry	52.3	44.1	47.4
Fish/sea food	9.5	8.4	8.9
Egg	25.7	23.6	24.5
Soybean	10.6	9.3	9.9
Fresh fruit	23.1	31.7	28.2
Dairy products	10.7	12.7	11.9
Spicy food	37.1	35.1	35.9
Food supplement (e.g. vitamins)	14.4	19.3	17.3
Exposure to domestic air pollution			
Use of smoky cooking fuel[b]	38.9	52.9	49.3
Home smoky in winter	40.6	39.5	40.0
Reproductive history in women			
Age at menarche <13 years		5.5	
Age at first live birth <20 years		9.4	
Having five or more live births		6.6	
Age at menopause <50[c]		52.0	
Ever used contraceptive pill		9.8	

[a]Apart from total and gender-specific variables, all other values in the table are adjusted for age.
[b]Restricted to participants who reported doing some cooking at home.
[c]Restricted to women who were aged $\geqslant$50 years at baseline survey.

Table 2 Anthropometric characteristics, physical activities and prevalence of prior disease at baseline survey, 2004–08

Characteristics[a]	Men ($n = 210\,222$), %	Women ($n = 302\,669$), %	Total ($n = 512\,891$), %
Height (cm)			
<155	4.6	55.7	34.7
155–159	14.9	28.3	22.8
160–164	27.9	13.0	19.1
⩾165	52.5	3.1	23.4
Mean (SD)	165.4 (6.5)	154.0 (6.0)	158.7 (8.3)
BMI (kg/m^2)			
<18.5	4.3	4.4	4.4
18.5 to <22.5	36.7	32.7	34.3
22.5 to <25	28.0	28.6	28.4
25 to <30	28.1	29.4	28.8
⩾30	2.9	4.9	4.1
Mean (SD)	23.4 (3.2)	23.8 (3.5)	23.7 (3.4)
Waist circumference (cm)			
<70	10.1	16.7	14.0
70–79	33.6	38.6	36.5
80–89	34.4	31.4	32.6
90–99	17.9	11.1	13.9
⩾100	3.9	2.3	2.9
Mean (SD)	82.0 (9.8)	79.1 (9.5)	80.3 (9.8)
Body fat percentage			
<15	13.1	0.4	5.6
15–24	56.0	15.1	31.9
25–34	28.6	51.7	42.3
⩾35	2.4	32.7	20.2
Mean (SD)	22.0 (6.2)	32.1 (7.1)	27.9 (8.4)
Systolic BP (mmHg)			
<100	1.9	4.4	3.4
100–119	24.8	31.0	28.4
120–139	43.5	36.4	39.3
140–159	20.2	17.9	18.9
⩾160	9.4	10.4	10.0
Mean (SD)	132 (20)	130 (22)	131 (21)
Blood glucose (mmol/l)			
<5.0	30.4	21.4	25.5
5.0–7.9	60.1	68.3	66.0
8.0–11.0	5.1	5.4	5.4
⩾11.1	2.8	3.1	3.1
Mean (SD)	5.9 (2.3)	6.2 (2.3)	6.1 (2.3)
Total physical activity (MET hours/day)			
<10	7.8	2.5	4.7
10–19.9	26.0	25.5	25.8
20–29.9	35.1	38.2	36.9
30–39.9	21.9	23.0	22.6
⩾40	9.2	10.7	10.1
Mean (SD)	25.5 (12.0)	26.6 (10.3)	26.1 (11.1)
Prior disease history			
IHD	2.5	3.4	3.0
Stroke/TIA	2.2	1.4	1.7
Diabetes	2.7	3.4	3.2
TB	2.0	1.2	1.5
Chronic respiratory diseases	3.0	2.3	2.6
Chronic hepatitis/cirrhosis	1.7	0.8	1.2
Peptic ulcer	5.3	2.9	3.9
Rheumatoid arthritis	1.4	2.5	2.1
Cancer	0.4	0.5	0.5

TIA, Transient Ischaemic Attack; TB: Tuberculosis.
[a]Adjusted for age.

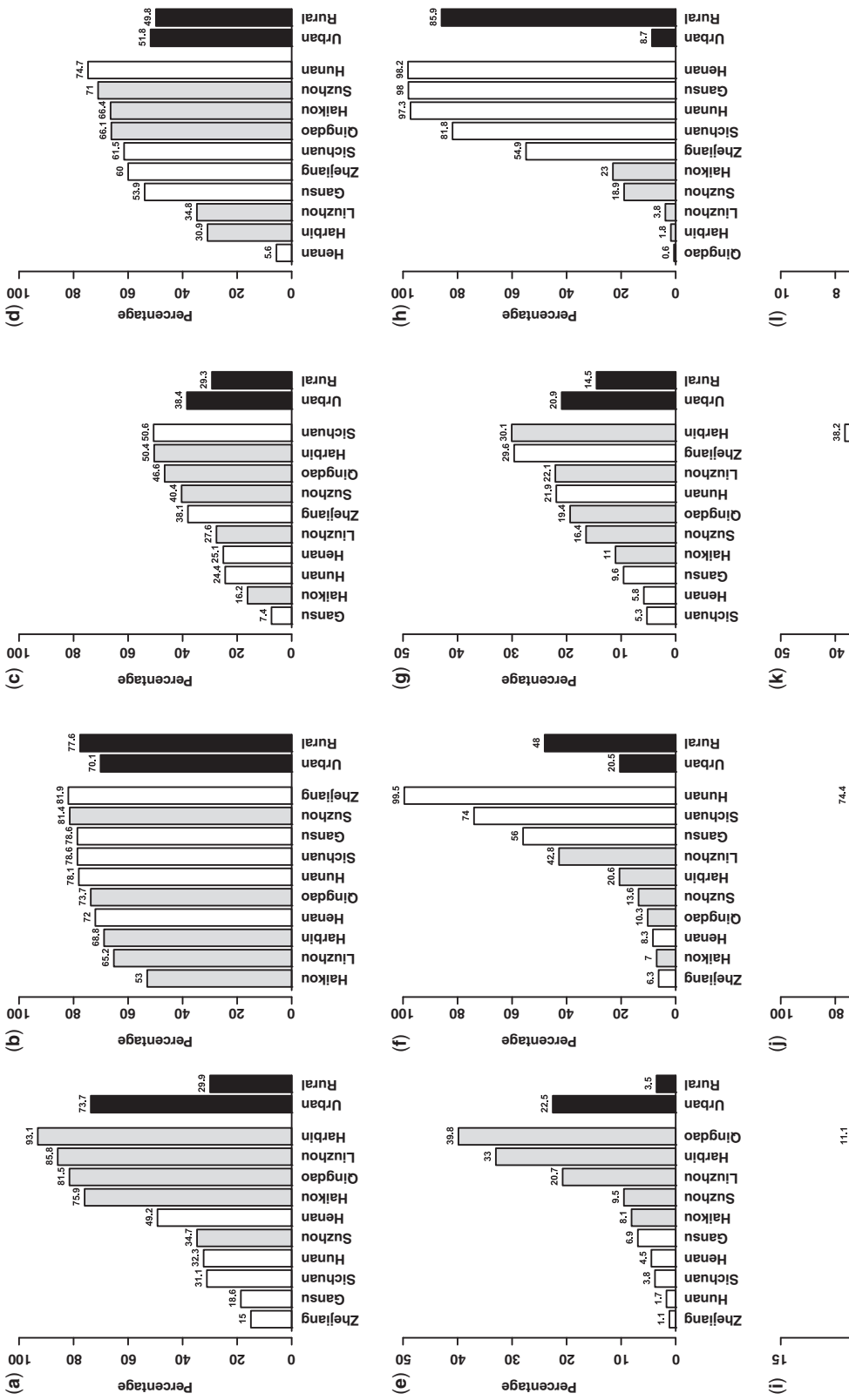

Figure 3 Variation of selected baseline variables across 10 study areas. Adjusted for age and for male and female combined results, additional adjustment was also made for sex. (a) Overall: per cent with $\geqslant$6 years education; (b) men: per cent ever-regular smokers, (c) men: per cent drinking alcohol weekly, (d) men: per cent drinking tea regularly; (e) overall: per cent consuming dairy food regularly; (f) overall: per cent eating spicy food regularly; (g) overall: per cent taking food supplements; (h) overall: per cent using smoky cooking fuel; (i) overall: per cent obese; (j) women (65–74 years): per cent with five or more live births; (k) women (35–44 years): per cent using oral contraceptives; and (l) overall: per cent with diabetes reported. Open bars indicate rural areas and grey bars indicate urban areas

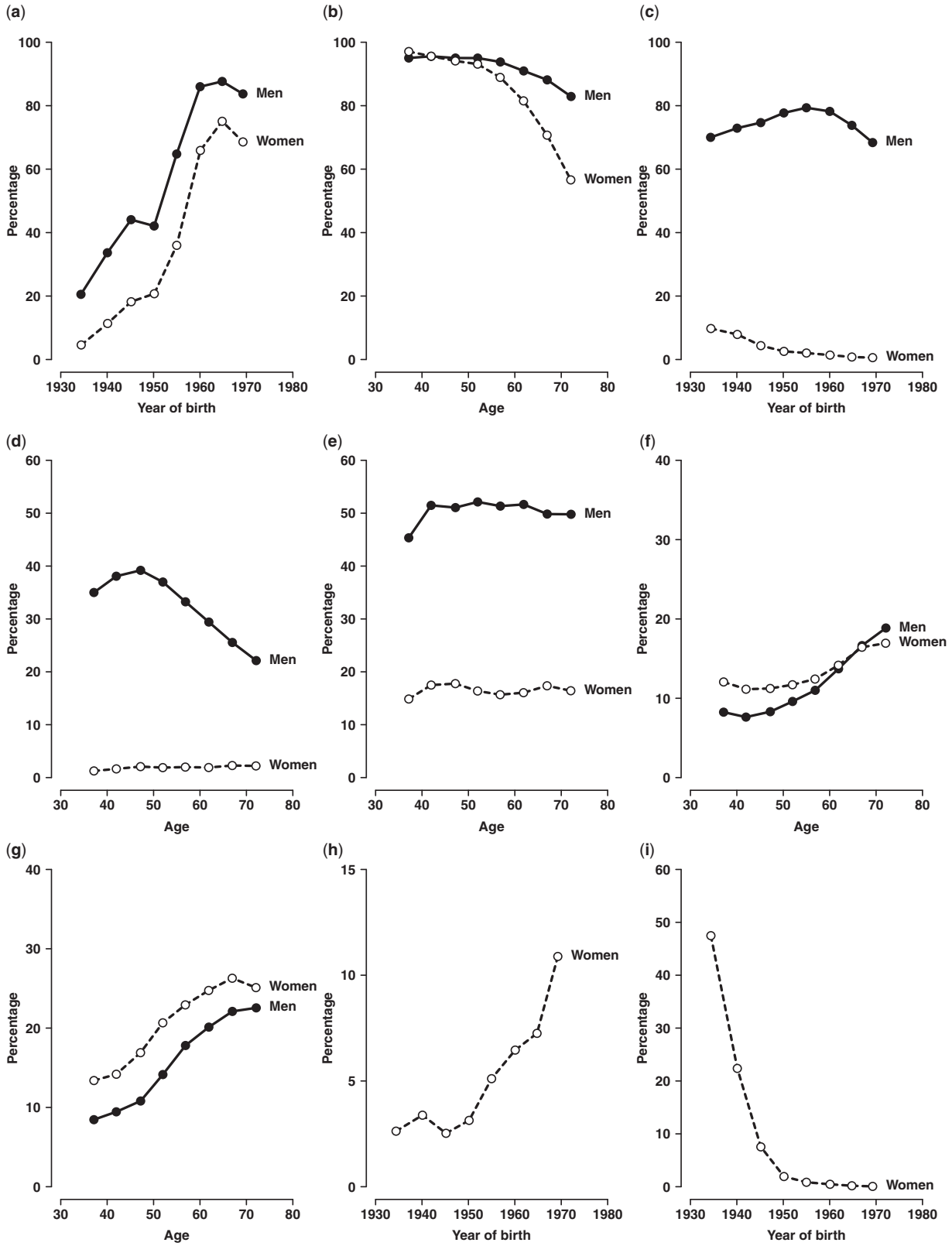

Figure 4 Prevalence of selected baseline variables by sex and by age or year of birth. Adjusted for area. (a) Overall: per cent with $\geqslant$6 years education; (b) overall: per cent married with spouse; (c) overall: per cent ever-regular smokers; (d) overall: per cent drinking alcohol weekly; (e) overall: per cent drinking tea regularly; (f) overall: per cent consuming dairy food regularly; (g) overall: per cent taking food supplements; (h) women: per cent with age at menarche $<$13 years; and (i) women: per cent with five or more live births

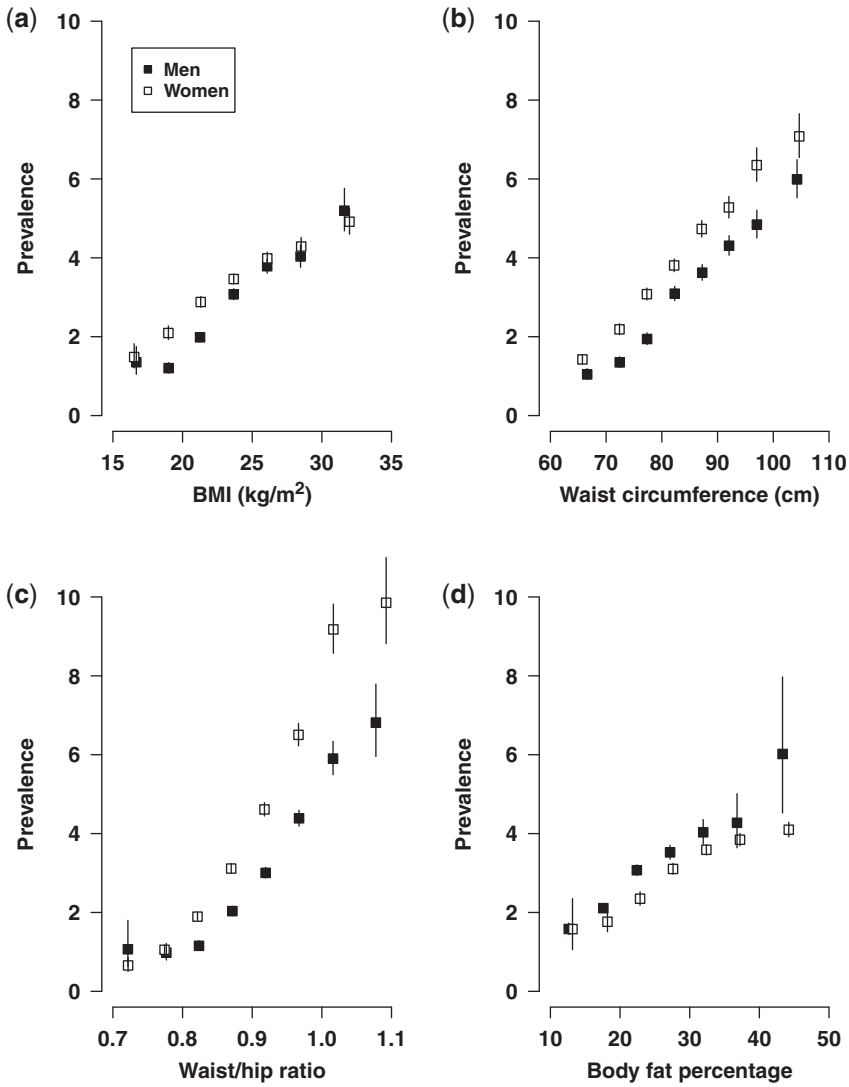

Figure 5 Associations of different measures of adiposity with self-reported diabetes at baseline. Prevalence of diabetes vs (a) BMI, (b) waist circumference, (c) waist/hip ratio, (d) body fat percentage after adjustment for age and area. Solid boxes denote men and open boxes denote women. Vertical line indicates 95% confidence interval (CI)

chronic conditions (Table 2) For certain conditions, such as diabetes, there was large variation across the 10 study areas (Figure 3) and inclusion of random blood glucose measured at baseline did not alter the patterns of regional variation. Waist/hip ratio was found to be more strongly associated with diabetes compared with other measures of adiposity in both men and women (Figure 5).

At the baseline, QC survey data were available for 15 728 participants (3.1%), with the mean length of time between baseline and QC survey being 17 days [standard deviation (SD) = 36 days]. There was good agreement between the baseline and QC survey for several common variables, with the weighted κ coefficient being 0.94 for tobacco use, 0.79 for alcohol drinking, 0.77 for tea drinking, 0.94 for women's menopausal status, 0.91 for number of pregnancies, 0.85 for age at menopause and 0.76 for use of contraceptive pills. The within-person Spearman correlation coefficients between baseline and QC measures were 0.84 for systolic blood pressure (SBP) and 0.77 for diastolic blood pressure (DBP). For the re-survey in 2008, of the total invited, 19 802 (80%) attended it, with the response rate being higher in rural (84%) than in urban areas (75%), and slightly higher in women (81%) than in men (78%). For most of the variables examined, especially various physical measurements (Table 3), there was good agreement with the baseline measures. The height, weight and BMI showed extremely high correlation with baseline measures (0.99, 0.96 and 0.93, respectively), whereas for other measures of adiposity (waist and hip circumferences, and body fat percentage), they ranged from 0.82 to 0.90. For SBP, it was 0.70 without any

Table 3 Spearman correlation coefficients for selected physical measurements between baseline survey and re-survey among 19 788 participants

Variables	Men ($n = 7770$)	Women ($n = 12\,018$)	Overall[a] ($n = 19\,788$)
Standing height	0.98	0.99	0.99
Sitting height	0.89	0.89	0.92
Weight	0.96	0.95	0.96
BMI	0.94	0.93	0.93
Waist circumference	0.87	0.82	0.84
Hip circumference	0.82	0.81	0.82
Body fat percentage	0.81	0.85	0.90
Heart rate	0.51	0.53	0.53
SBP	0.65	0.72	0.70
DBP	0.64	0.67	0.66

[a]Not adjusted for sex. For sitting height and body fat percentage, the sex-adjusted correlation coefficients are 0.88 and 0.84, respectively, whereas for other variables little change was seen.

adjustment for months of measurement, which affected blood pressure greatly (and heart rate to a lesser extent) due to seasonal variation in outdoor temperature (data not shown).

By 1 January 2011, 10 763 people (2.1%) were known to have died, 956 (0.2%) were lost to follow-up (Table 4). Overall, linkage to local HI databases had already been achieved for 91% of the participants, ranging from 76% in Haikou to 99% in Sichuan. Based on death and disease registries (but not HI), there were 9475 new cases of stroke (including 8008 first incident events), 4071 IHD and 6381 cancer cases reported.

Discussion

The CKB is one of the largest blood-based prospective studies ever conducted. It aims to assess the effects of both established and emerging risk factors for many different diseases, not only overall but also in a range of different circumstances (e.g. at different levels of other risk factors). By storing both plasma and DNA samples, it will also allow reliable assessment of the relevance of many genetic and other novel blood-related factors that will be proposed in the future as determinants of chronic diseases. To achieve rapid recruitment cost-effectively and to target regions with high rates of certain conditions (e.g. stroke, chronic obstructive pulmonary disease [COPD]), the study cohort is not designed to be representative of the general population in China. Despite this, the inclusion of an extremely large number of people from diverse populations should help to generate important new findings about the causes of many diseases that will be generalizable to other populations with different distribution of risk exposures.[34] The collection of non-fatal disease events will not only greatly increase the statistical power of the study, but also improve the reliability of diagnosis for these conditions [e.g. ~80% of the reported stroke cases can be confirmed and subtyped based on computerised tomography (CT) or magnetic resonance imaging (MRI) scans]. The establishment of electronic linkage with the HI databases in the study areas, which was not envisaged at the beginning of the study,[27] will further increase the statistical power and the range of conditions that can be investigated reliably.

Given the extensive range of data collected, it is not feasible to provide detailed descriptive analyses in this article of all the information that was collected. Nevertheless, the results that are presented demonstrate great heterogeneity of many major risk factors for chronic disease within the study population. Of particular concern is the high prevalence of tobacco smoking in Chinese men, which has followed a similar pattern, albeit 40 years later, to that observed among adults in the USA.[15] On the whole, only a small proportion of Chinese women smoke and, unexpectedly, there has been a progressive decrease over the past few decades in the probability of women starting to smoke. If this low uptake of smoking by young women continues, then although the proportion of deaths before the age of 70 years that is attributed to smoking may increase from ~12% in 1990 to ~33% in 2030 among Chinese men, it will probably decrease from ~3 to <1% among Chinese women.[20,26,35] Long-term continuation of the present study will help monitor the tobacco epidemic over the next few decades in China.

For each of the other selected variables that are presented, there is also large heterogeneity by age, sex and study area. For some, the observed variation may be driven mainly by environmental factors (e.g. seasonality of blood pressure) or due to upheavals in recent Chinese history (e.g. number of participants by birth year and anomalous changes in secular trend of age at menarche); for some, it may be due to birth cohort effects (e.g. age at menarche and birth

Table 4 Status of long-term follow-up for mortality and hospitalized events in 10 study areas by 1 January 2011

Study area	Number of participants	Number died	Number lost	Percentage linked to HI database[a]	Year when HI started[b]
Qingdao	35 509	358	117	97	2002
Harbin	57 555	1146	497	87	2001
Haikou	29 689	99	0	76	2005
Suzhou	53 260	813	101	91	2004
Liuzhou	50 173	850	0	89	2000
Sichuan	55 687	1682	25	99	2006
Gansu	50 041	1077	17	91	2006
Henan	63 357	1718	11	89	2005
Zhejiang	57 704	1327	13	97	2002
Hunan	59 916	1693	175	90	2006
Total	512 891	10 763	956	91	

[a]Linkage with HI databases was achieved through the unique national ID number and further checks and linkage using other procedures will be done for unmatched participants.
[b]In most urban areas, the good coverage of HI in general population was achieved mainly after 2005.

rates at different ages in women) or due to different stages of socio-economic development (e.g. prevalence of obesity and of diabetes across different areas); for some it may be accounted for by local traditions that have persisted for generations (e.g. eating spicy food in certain regions) or due to age-related changes in certain dietary patterns (e.g. eating dairy food and taking food supplements) and for many it could well involve a combination of different factors. Understanding these variations by age and sex, or by geographic location, is not the primary aim of the present prospective cohort study, but the large heterogeneity observed for most of the variables studied (probably much more extreme than in many other populations) will greatly increase the scientific value of the study. For example, with many low values for BMI (and probably for blood cholesterol), both the risks and benefits associated with low BMI (or low cholesterol) can be assessed, avoiding confusion between what is 'statistically' normal (e.g. average BMI of $\sim$30 kg/m^2 in the USA)[36] and what is 'biologically' optimal (e.g. average BMI of 20–22 kg/m^2 among most of the rural participants in the present study), which is associated with the lowest overall mortality.[14] Such prospective evidence will be essential for assessing the appropriateness of having regional-specific cut-points for defining overweight and obesity.[18] Moreover, the availability of different measures of adiposity (e.g. BMI, waist circumference and body fat percentage) in the present study will also allow for reliable assessment of their relative values for different conditions. For example, waist/hip ratio was shown to be more strongly associated with prevalence of diabetes compared with other measures of adiposity, but this needs to be confirmed further by prospective analysis of the disease incidence data. The inclusion of large number of people born during the great famine in 1959–61 also provides a good opportunity to assess the relevance of nutritional deprivation early in life to chronic diseases later in life.[37]

In summary, we have successfully established the large CKB and good linkages with various health record systems for long-term follow-up of study participants. With another 5 years of follow-up, there will be about 25 000 deaths and about 100 000 hospitalized events. This will help to provide reliable evidence about the effects of smoking, adiposity, blood pressure and many risk factors for major diseases. Stroke is the largest cause of serious disability and death in China.[38,39] With accumulation of large number of well-characterized stroke cases among the study participants, it will soon allow a uniquely large blood-based nested case–control study of genetic and non-genetic causes of stroke to be conducted. As follow-up continues, subsequent studies of a wide range of risk factors for a range of other common conditions will also be possible. This large biobank will be a powerful resource for investigating, both independently and in collaboration with other similar studies around the world,[40–43] the main causes of many common chronic diseases over the next few decades. The information generated will be of general relevance to the better understanding of disease aetiology not only in China but also in other countries.

Funding

The baseline survey and first re-survey in China were supported by a research grant from the Kadoorie Charitable Foundation in Hong Kong; follow-up of the project during 2009–14 is supported by the Wellcome Trust in the UK (grant 088158/Z/09/Z);

the Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU) at Oxford University also receives core funding for it from the UK Medical Research Council, the British Heart Foundation, and Cancer Research UK.

Acknowledgements

We thank Judith MacKay in Hong Kong; Yu Wang, Gonghuan Yang, Zhengfu Qiang, Lin Feng, Maigen Zhou, Wenhua Zhao and Yan Zhang in China CDC; Lingzhi Kong, Xiucheng Yu and Kun Li in the Chinese Ministry of Health; and Yiping Chen, Sarah Clark, Martin Radley, Hongchao Pan, Jill Boreham, Gary Whitlock, Paul Sherliker and Sarah Lewington in the CTSU, Oxford, for assisting with the design, planning, organization, conduct of the study and data analysis. The most important acknowledgement is to the participants in the study and the members of the survey teams in each of the 10 regional centres, as well as to the project development and management teams based at Beijing, Oxford and the 10 regional centres. CTSU acknowledges support from the BHF Center of Research Excellence, Oxford.

Members of China Kadoorie Biobank collaborative group

(a) International Steering Committee

Liming Li, Zhengming Chen, Junshi Chen, Rory Collins, Fan Wu (ex-member), Richard Peto.

(b) Study coordinating centres

International (ICC, Oxford): Zhengming Chen, Garry Lancaster, Xiaoming Yang, Alex Williams, Margaret Smith, Ling Yang, Yumei Chang

National (NCC, Beijing): Yu Guo, Guoqing Zhao, Zheng Bian, Lixue Wu, Can Hou

Regional (RCC, 10 areas in China):

Qingdao

Qingdao CDC: Zengchang Pang, Shaojie Wang, Yun Zhang, Kui Zhang

Licang CDC: Silu Liu

Heilongjiang

Provincial CDC: Zhonghou Zhao, Shumei Liu, Zhigang Pang

Nangang CDC: Weijia Feng, Shuling Wu, Liqiu Yang, Huili Han, Hui He

Hainan

Provincial CDC: Xianhai Pan, Shanqing Wang, Hongmei Wang

Meilan CDC: Xinhua Hao, Chunxing Chen, Shuxiong Lin

Jiangsu

Provincial CDC: Xiaoshu Hu, Minghao Zhou, Ming Wu,

Suzhou CDC: Yeyuan Wang, Yihe Hu, Liangcai Ma, Renxian Zhou, Guanqun Xu

Guanxi

Provincial CDC: Baiqing Dong, Naying Chen, Ying Huang

Liuzhou CDC: Mingqiang Li, Jinhuai Meng, Zhigao Gan, Jiujiu Xu, Yun Liu

Sichuan

Provincial CDC: Xianping Wu, Yali Gao, Ningmei Zhang

Pengzhou CDC: Guojin Luo, Xiangsan Que, Xiaofang Chen

Gansu

Provincial CDC: Pengfei Ge, Jian He, Xiaolan Ren

Maiji CDC: Hui Zhang, Enke Mao, Guanzhong Li, Zhongxiao Li, Jun He

Henan

Provincial CDC: Guohua Liu, Baoyu Zhu, Gang Zhou, Shixian Feng

Huixian CDC: Yulian Gao, Tianyou He, Li Jiang, Jianhua Qin, Huarong Sun

Zhejiang

Provincial CDC: Liqun Liu, Min Yu, Yaping Chen

Tongxiang CDC: Zhixiang Hu, Jianjin Hu, Yijian Qian, Zhiying Wu, Lingli Chen

Hunan

Provincial CDC: Wen Liu, Guangchun Li, Huilin Liu

Liuyang CDC: Xiangquan Long, Youping Xiong, Zhongwen Tan, Xuqiu Xie, Yunfang Peng

Conflict of interest: None declared.

KEY MESSAGES

- A total of 512 891 men and women aged 30–79 years were recruited from 10 geographically diverse areas of China, with extensive data collection and long-term storage of blood samples.
- For each of the main baseline variables analysed (e.g. smoking, alcohol, BMI, blood pressure and prior history of diabetes), there is large heterogeneity by age, sex and study area.
- The established linkages with mortality and morbidity registries as well as with national HI system will soon allow for reliable prospective assessment of main genetic and non-genetic determinants of a range of common conditions in Chinese population.

References

[1] Murray CJ, Lopez AD. Global mortality, disability, and the contribution of risk factors: Global Burden of Disease Study. *Lancet* 1997;**349:**1436–42.

[2] Yusuf S, Reddy S, Ounpuu S, Anand S. Global burden of cardiovascular diseases: part I: general considerations, the epidemiologic transition, risk factors, and impact of urbanization. *Circulation* 2001;**104:**2746–53.

[3] Yang GH, Murray CJL, Zhang Z (eds). *Exploring Adult Mortality in China: Levels, Patterns and Causes*. Beijing: Hua Xia Press, 1991.

[4] Wang L, Kong L, Wu F, Bai Y, Burton R. Preventing chronic diseases in China. *Lancet* 2005;**366:**1821–24.

[5] Yang G, Kong L, Zhao W *et al*. Emergence of chronic non-communicable diseases in China. *Lancet* 2008;**372:** 1697–705.

[6] Yang G, Fan L, Tan J *et al*. Smoking in China: findings of the 1996 National Prevalence Survey. *JAMA* 1999;**282:** 1247–53.

[7] Wang H, Du S, Zhai F, Popkin BM. Trends in the distribution of body mass index among Chinese adults, aged 20-45 years (1989-2000). *Int J Obes* 2007;**31:**272–28.

[8] Li JY, Liu BQ, Li GY, Chen ZJ, Sun XI, Rong SD. Atlas of cancer mortality in the People's Republic of China. An aid for cancer control and research. *Int J Epidemiol* 1981;**10:** 127–33.

[9] Chen JS, Campbell TC, Li JY, Peto R (eds). *Diet, Lifestyle and Mortality in China: A Study of the Characteristics of 65 Chinese Counties*. Oxford: Oxford University Press, 1990.

[10] Chen JS, Peto R, Pan WH, Liu B, Campbell TC (eds). *Mortality, Biochemistry, Diet and Lifestyle in Rural China: Geographic Study of the Characteristics of 69 Counties in Mainland China and 16 Areas in Taiwan*. Oxford: Oxford University of Press, 2006.

[11] Collins FS, Green ED, Guttmacher AE, Guyer MS. A vision for the future of genomics research. *Nature* 2003; **422:**835–47.

[12] Davey Smith G, Ebrahim S, Lewis S, Hansell AL, Palmer LJ, Burton PR. Genetic epidemiology and public health: hope, hype, and future prospects. *Lancet* 2005;**366:** 1484–98.

[13] Hunter DJ. Gene-environment interactions in human diseases. *Nat Rev Genet* 2005;**6:**287–98.

[14] Prospective Studies Collaboration. Body-mass index and cause-specific mortality in 900 000 adults: collaborative analyses of 57 prospective studies. *Lancet* 2009;**373:** 1083–96.

[15] Peto R, Lopez AD, Boreham J, Thun M, Heath C (eds). *Mortality from Smoking in Developed Countries 1950-2000: Indirect Estimates from National Vital Statistics*. Oxford: Oxford University Press, 1994.

[16] Prospective Studies Collaboration. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *Lancet* 2002;**360:**1903–13.

[17] Prospective Studies Collaboration. Blood cholesterol and vascular mortality by age, sex, and blood pressure: a meta-analysis of individual data from 61 prospective studies with 55,000 vascular deaths. *Lancet* 2007;**370:** 1829–39.

[18] WHO Expert Consultation. Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies. *Lancet* 2004;**363:**157–63.

[19] Chen ZM, Peto R, Collins R, MacMahon S, Lu J, Li W. Serum cholesterol concentration and coronary heart disease in population with low cholesterol concentrations. *BMJ* 1991;**303:**276–82.

[20] Niu SR, Yang GH, Chen ZM *et al*. Emerging tobacco hazards in China: 2. Early mortality results from a prospective study. *BMJ* 1998;**317:**1423–24.

[21] He J, Gu D, Wu X *et al*. Major causes of death among men and women in China. *N Engl J Med* 2005;**353:** 1124–34.

[22] Yuan JM, Ross RK, Gao YT, Yu MC. Body weight and mortality: a prospective evaluation in a cohort of middle-aged men in Shanghai, China. *Int J Epidemiol* 1998;**27:**824–32.

[23] Zheng W, Chow WH, Yang G *et al*. The Shanghai Women's Health Study: rationale, study design, and baseline characteristics. *Am J Epidemiol* 2005;**162:** 1123–31.

[24] Jiang C, Thomas GN, Lam TH *et al*. Cohort profile: The Guangzhou Biobank Cohort Study, a Guangzhou-Hong Kong-Birmingham collaboration. *Int J Epidemiol* 2006;**35:** 844–52.

[25] Sai XY, He Y, Men K *et al*. All-cause mortality and risk factors in a cohort of retired military male veterans, Xi'an, China: an 18-year follow up study. *BMC Public Health* 2007;**7:**290.

[26] Liu BQ, Peto R, Chen ZM *et al*. Emerging tobacco hazards in China: 1. Retrospective proportional mortality study of one million deaths. *BMJ* 1998;**317:**1411–22.

[27] Chen ZM, Lee LM, Chen JS *et al*. Cohort profile: the Kadoorie Study of Chronic Disease in China (KSCDC). *Int J Epidemiol* 2005;**34:**1243–49.

[28] Kessler RC, Andrews G, Mroczec C. The World Health Organization Composite International Diagnostic Interview Short-Form (CIDI-SF). *Int J Methods Psychiatr Res* 1998;**7:**171–85.

[29] Clarke R, Shipley M, Lewington S *et al*. Underestimation of risk associations due to regression dilution in long-term follow-up of prospective studies. *Am J Epidemiol* 1999;**150:**341–53.

[30] Kaaks R, Riboli E. Validation and calibration of dietary intake measurements in the EPIC project: methodological considerations. European Prospective Investigation into Cancer and Nutrition. *Int J Epidemiol* 1997;**26(Suppl 1):** S15–25.

[31] Yang G, Rao C, Ma J *et al*. Validation of verbal autopsy procedures for adult deaths in China. *Int J Epidemiol* 2006; **35:**741–48.

[32] Ainsworth BE, Haskell WL, Leon AS *et al*. Compendium of physical activities: classification of energy costs of human physical activities. *Med Sci Sports Exerc* 1993;**25:** 71–80.

[33] Ainsworth BE, Haskell WL, Whitt MC *et al*. Compendium of physical activities: an update of activity codes and MET intensities. *Med Sci Sports Exerc* 2000;**32(Suppl 9):** S498–504.

[34] Manolio TA, Collins R. Enhancing the feasibility of large cohort studies. *JAMA* 2010;**304:**2290–91.

[35] Peto R, Chen ZM, Boreham J. Tobacco: The growing epidemic in China. *CVD Prevent Control* 2009;**4:**61–70.

[36] WHO Global InfoBase team. *Surveillance of Chronic Disease Risk Factors: Country-level Data and Comparable Estimates*. Geneva: World Health Organization, 2005.

[37] Barker DJ. Fetal origins of coronary heart disease. *BMJ* 1995;**311:**171–74.
[38] Wu Z, Yao C, Zhao D *et al*. Sino-MONICA project: a collaborative study on trends and determinants in cardiovascular diseases in China, Part I: morbidity and mortality monitoring. *Circulation* 2001;**103:**462–68.
[39] Zhao D, Liu J, Wang W *et al*. Epidemiological transition of stroke in China: twenty-one-year observational study from the Sino-MONICA-Beijing Project. *Stroke* 2008;**39:** 1668–74.
[40] Riboli E, Kaaks R. The EPIC Project: rationale and study design. European prospective investigation into cancer and nutrition. *Int J Epidemiol* 1997;**26(Suppl 1):** S6–14.
[41] Tapia-Conyer R, Kuri-Morales P, Alegre-Diaz J *et al*. Cohort profile: the Mexico City Prospective Study. *Int J Epidemiol* 2006;**35:**243–49.
[42] UK Biobank. 2010. http://www.ukbiobank.ac.uk/ (30 June 2011, date last accessed).
[43] Fortier I, Burton PR, Robson PJ *et al*. Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. *Int J Epidemiol* 2010;**39:** 1383–93.