

Cohort Profile: The Korean Cancer Prevention Study-II (KCPS-II) Biobank

Yon Ho Jee^{1†}, Jonathan Emberson^{2†}, Keum Ji Jung³, Sun Ju Lee³, Sunmi Lee⁴,
Joung Hwan Back⁴, Seri Hong⁵, Heejin Kimm^{3,5}, Paul Sherliker², Sun Ha Jee^{3,5††},
Sarah Lewington^{2††}

¹ Nuffield Department of Population Health, University of Oxford, UK

² MRC Population Health Research Unit, Clinical Trial Service Unit and
Epidemiological Studies Unit, Nuffield Department of Population Health, University of
Oxford, UK

³ Institute for Health Promotion, Graduate School of Public Health, Yonsei University,
Seoul, Korea

⁴ Health Insurance Policy Research Institute, National Health Insurance Service,
Wonju, Korea

⁵ Department of Epidemiology and Health Promotion, Graduate School of Public
Health, Yonsei University, Seoul, Korea

Corresponding author:

Sun Ha Jee, PhD, MPH

Department of Epidemiology and Health Promotion,
Graduate School of Public Health, Yonsei University,
50-1 Yonse-ro, Seodaemun-gu, Seoul, Korea.

Tel:+82-2-2228-1523

Fax: 82-2-365-5118

E-mail: jsunha@yuhs.ac

Word count: 2167 main text. 3 tables, 3 figures

Why was the cohort set up?

The Korean Cancer Prevention Study-II (KCPS-II) Biobank is a large blood-based cohort study with long term follow-up via a unique linkage of routine, nationwide medical examinations conducted at health promotion centres across South Korea with records for mortality and hospitalization. It was initiated in April 2004 and has been supported by the Seoul city government since December 2005. In addition to the examination of cancer, a major reason for the cohort being established was to examine the determinants and long-term consequences of the metabolic syndrome – a term used to describe the co-occurrence of increased blood pressure, high blood sugar, excess central adiposity, and elevated cholesterol or triglyceride levels – which has been shown in many studies to be associated with a substantially increased risk of cardiovascular diseases (CVD),¹⁻⁴ as well as some cancers.^{5, 6} However, through its large sample size, detailed mortality and morbidity linkage, and, crucially, availability of baseline blood samples for all participants, the cohort is now well-placed to assess the relevance to health of a much wider range of personal health behaviours and physical and biological characteristics (including genetics).

Who is in the cohort?

The cohort comprises 156,701 participants (94 840 men and 61 861 women) who undertook routine health assessments, provided blood samples and gave informed consent for long-term prospective follow-up. Recruitment was through eighteen health promotion centres across South Korea (**Figure 1**), starting in 2004 at three centres (Severance Hospital [2 centres], and Bundang Cha Hospital), before being expanded in April 2006 to a further fifteen centres (Korean Medical Institute [7 centres across South Korea], Ewha Womans University Mokdong Hospital, Bundang Seoul National University Hospital, Korea University Hospital, Kyung Hee University Hospital, The Catholic University of Korea Buchun St. Marys Hospital, Seoul Medical Center, Asan Medical Center, and Hanyang University Guri Hospital). Median recruitment was achieved by the end of 2007 with 90% of participants recruited between mid-2005 and the end of 2008 (**Figure 2**). 90% of participants were recruited from the Seoul and Gyeonggi provinces (whose combined population of about 19 million people represents 40% of South Korea's population).⁷ The main

reason given for participants visiting the centres were: regular check-up (33%); health issue (29%); mandatory group examination from work (16%); interest in current health condition (10%); referral by the family and/or relatives (6%); referral by a colleague or friends (3%); referral by a doctor (2%); and other reasons (2%). Participants agreed to join the study after reading and agreeing a summary of the research plan, and the response rate was over 90%. The study was approved by the Institutional Review Board of Human Research of Yonsei University as well as all of the participating health promotion centres.

What has been measured?

Health assessment:

Appointments were scheduled for between 7am and 9am each day; the data collected are summarized in **Table 1**. The health assessment involved self-completion of a questionnaire which contained questions about health status and associated behaviours; attendees then took part in a medical examination. Although there were some differences in the study questionnaires and equipment used by individual study centres, a core set of both were utilized. Self-reported items included information on social status (insurance premium was recorded for all participants and used as a proxy for social class, but in 6 of the 18 centres additional information was collected on educational attainment, marital status, occupation, and income), medical history, medication use, and behavioural risk factors (alcohol, smoking, and physical activity). Weight, height and waist circumference were recorded while participants wore light clothing. Seated blood pressure after a 5-minute rest was measured by a registered nurse or blood pressure technician. Most centres used standard mercury sphygmomanometers, but some used automatic sphygmomanometers. In 9 of the centres, pulmonary function was also measured.

Blood sampling and assays:

For their morning appointment, participants were asked to fast overnight. EDTA blood and serum was taken and stored at -70C for future study. In addition, EDTA blood was drawn for immediate use to measure a range of biomarkers (Table 1). Glucose, total cholesterol, triglyceride, high density lipoprotein-cholesterol (HDL-c),

low density lipoprotein-cholesterol (LDL-c), and other biomarkers were measured in the hospital laboratory by a COBAS INTEGRA 800 and a 7600 Analyzer (Hitachi, Tokyo, Japan). LDL-c was either directly measured or calculated based on total, HDL-c and triglyceride levels. Each laboratory had internal and external quality control procedures as required by the Korean Association of Laboratory Quality Control. Agreement for each biomedical marker across individual hospitals was high (correlation 0.96–0.99).

How often have they been followed up?

Re-surveys

Participants may have attended for re-examinations, similar to the baseline assessment, every year. By the end of the baseline recruitment, 71 580 participants (46%) had had at least one such repeat examination (on average the first repeat visit was just over a year after their baseline assessment), 20 694 (13%) had had at least two repeat examinations, and 7033 (5%) had had at least three repeat examinations. At each examination, participants complete a questionnaire and provide a fasting blood sample (with the same procedures as at recruitment). These repeat examinations will allow future prospective analyses to correct for the regression dilution bias caused by within-person variability in characteristics recorded at recruitment.⁸

Follow-up for cancer incidence, hospital admissions and cause-specific mortality

Every South Korean national is assigned a personal identification number at birth, and this number allows linkage with the national cancer centre (NCC) registry, hospital admission records, and death registers. Moreover, all Koreans are members of the National Health Insurance Service (NHIS) (formerly, the Korean Medical Insurance Corporation and National Health Insurance Corporation) allowing linkage with all hospital admissions. To ensure anonymity, all linkages are carried out by NCC and NHIS staff. Using computerized searches of data provided by the National Cancer Center, National Statistical Office, and NHIS in Korea, outcomes are ascertained from the national cancer registry, diagnosis on hospital discharge summaries and from causes listed on death certificates. With emigration from South

Korea being very rare, it is anticipated that follow-up will be close to 100% complete.

The principal outcomes for future epidemiological analyses will be: (1) cause-specific mortality; (2) morbidity and mortality from ischaemic heart disease, stroke, and other vascular diseases, as well as from cancer and other major causes of morbidity. During an average of 8 years' follow-up (to December 2015) 1184 participants had died, 7252 had had an incident vascular event, 5669 an incident cancer and 6349 an incident diagnosis of diabetes (**Table 2**).

What has it found? Key findings and publications

Baseline characteristics

Table 3 shows the characteristics at recruitment of the 94 840 men and 61 861 women. Mean age at recruitment was 42 (SD 11) years and mean BMI was 23.6 (SD 3.0) kg/m² (slightly higher in men, 24.4 [3.0], than women, 22.2 [3.0]). Consequently, 40% of men but only 17% of women had BMI >25kg/m² (ie, were overweight or obese). Overall, 3.6% of participants reported having been previously diagnosed with diabetes or were taking treatment for diabetes (4.3% of men, 2.4% of women), and mean (SD) fasting glucose concentration among such participants was 7.7 (2.6) mmol/L (similar among men and women). However, an additional 2.1% of men and 1.0% of women had fasting glucose >7mmol/L so that, overall, 6.4% of men and 3.4% of women had previously diagnosed/treated diabetes or a fasting glucose >7 mmol/L. Mean (SD) systolic blood pressure (SBP) and diastolic blood pressure (DBP) at recruitment were 118 (14) and 74 (10) mmHg, respectively, and were somewhat lower in women (112/71 mmHg) than men (121/77 mmHg). While cigarette smoking was common in men, with over two-thirds of them reporting ever having smoked (44% current and 27% former), it was very uncommon in women (4% current, 4% former). 42% of participants reported doing no regular recreational physical activity, slightly higher in women than men (54% vs 34% respectively).

Figure 3 shows the sex-specific trends in vascular risk factors either by birth cohort or by age. The most notable secular trends were in smoking habits and alcohol use. Among men born before 1970, 69% had ever smoked, and this was fairly constant,

but a lower proportion of men born after 1970 reported having ever smoked. However, older men were much more likely to have quit than younger men, so current smoking was more than twice as common in younger men: 49% among men born in the 1970s, compared with 30% for those born in the 1950s. Very few women reportedly smoked in any birth cohort. Overall, 59% of men reported drinking alcohol regularly (at least weekly), and this increased only slightly with age from about 50% in men born ~1945 to 60% in men born ~1980. By contrast regular (at least weekly) alcohol drinking has increased markedly among the women from 7% in women born ~1945 to nearly 40% in women born ~1980. The prevalence of obesity was low overall, and was slightly higher in younger than older men, but lower in younger than older women. The prevalence of diabetes (self-reported, treated, or fasting glucose >7mmol/L), however, increased markedly with age in both men and women from <5% before age 40 to about 25% by age 70. In men, both SBP and DBP rose only slightly with age from 121/75 mmHg among the youngest (average year of birth, 1980) to 126/78 mmHg among the oldest (average year of birth, 1950) whereas the increase was somewhat steeper in women (from 107/61 among the youngest women to 125/76 mmHg among the oldest).

Previous prospective analyses

KCPS-II Biobank data, or subsets of the data, have already been used in several epidemiological^{9, 10} and genetic studies,¹¹⁻¹⁴ and have contributed to international collaborative research projects.^{15, 16} These studies have shown that genetic variants in *CDH13* influence adiponectin levels in Korean adults¹¹ and that serum adiponectin is associated with a family history of diabetes independently of obesity and insulin resistance.⁹ We have also shown that genetic variants in *SLC2A9* influence uric acid levels¹³ and contributed to a meta-analysis that identified four novel loci for waist-hip ratio and waist circumference.¹⁴ We have published two reports on colorectal cancer showing that serum glucose levels may be a potential marker for the disease¹⁰ and that 7 out of 23 genetic variants previously identified in populations of European descent may be useful for predicting risk of colorectal cancer in Koreans.¹²

Future analysis plans

As the cohort continues to mature, further analyses exploring the importance to health of a range of lifestyle and biological characteristics will be performed. This will include studying not only the metabolic syndrome and other traditional vascular and neoplastic risk factors, but also genetic factors as well as a wide range of other factors measurable in blood (eg, using high-throughput “omics” technologies, such as metabolomics). The cost-effectiveness of such analyses can be improved by employing case-cohort designs, and with this in mind we have established a randomly selected sub-cohort of 8000 subjects from the 156 701 study participants, which has been genotyped using an 830,000 Korean-specific SNP chip panel developed by the Korean Centre for Disease Control and Prevention and the Korean National Institute of Health.

What are the main strengths and weaknesses?

The main strengths of this study are its size, its reliable and detailed linkage to health records, and its collection of blood samples at recruitment in all participants. Since all participants have a unique identification number assigned at birth, follow-up is almost 100% complete, while cancer diagnoses are based on histological type, resulting in high accuracy. A feature of the study population is its relatively young age (the average baseline age was 42 years), meaning it will be able to evaluate the long-term effects of health behaviours and habits in early middle-age. However, this does also mean that it will be some time before sufficient numbers of major health events (such as vascular disease or cancer) occur before reliable statistical analyses can be done.

Some collected information (in particular on socioeconomic status and pulmonary function) was limited to just a few centres, so analyses that require such measures will only be possible in a subset of the overall cohort. In addition, the cohort is not representative of the entire South Korean population as study participants were recruited mainly from the Seoul and Gyeonggi regions. It will not, therefore, necessarily provide generalizable estimates of the prevalences of particular characteristics. However, by including a large number of people who differ according

to their lifestyles, their biological and their genetic characteristics, the study is very well-placed to deliver on its main aim of studying the associations between genetic- or environmental risk factors and future disease incidence.¹⁷

Can I get hold of the data? Where can I find out more?

The KCPS-II Biobank data are not freely available, but the study group has collaborated with several other groups to share study data and encourage new collaborations. Potential collaborators are invited to contact the Secretary General (Lee SJ) at the administrative office of the KCPS-II Biobank at the Institute for Health Promotion, Graduate School of Public Health, Yonsei University, Seoul, Korea.

Profile in a nutshell:

- The Korean Cancer Prevention Study-II (KCPS-II) Biobank is a large blood-based cohort study with long term follow-up via a unique linkage of routine, nationwide medical examinations conducted at health promotion centres across South Korea with records for mortality and hospitalization.
- Recruitment was through eighteen health promotion centres across South Korea, starting in 2004, with 90% of participants recruited between mid-2005 and the end of 2008. The study comprises 156,701 adults (94 840 men and 61 861 women). Mean age at recruitment was 42 years.
- Every South Korean national is assigned a personal identification number at birth, and this number allows linkage with the national cancer centre (NCC) registry, hospital admission records, and death registers. Emigration from South Korea is very rare, so it is anticipated that follow-up will be close to 100% complete.
- Participants may attend for re-examinations, similar to the baseline assessment, every year: 71 580 participants (46%) have had at least one such repeat examination (on average the first repeat visit was just over a year after their baseline assessment).
- The health assessment involved self-completion of a questionnaire which contained questions about health status and associated behaviours; attendees then took part in a medical examination during which weight, height, waist circumference and blood pressure were measured. For their morning appointment, participants were asked to fast overnight, and EDTA blood and serum was taken and stored at -70C for future study; EDTA blood was also drawn for immediate use to measure a range of biomarkers.
- The KCPS-II Biobank data are not freely available, but the study group has collaborated with several other groups to share study data and encourage new collaborations. Potential collaborators are invited to contact the Secretary General (Lee SJ) at the administrative office of the KCPS-II Biobank at the Institute for Health Promotion, Graduate School of Public Health, Yonsei University, Seoul, Korea.

Funding

This study was funded by the Seoul R&BD Program (10526) at baseline. Follow-up is funded by a grant of the Korean Health Technology R&D Project (HI14C2686) and a grant from the National R&D Program for Cancer Control (1631020), Ministry of Health & Welfare, Republic of Korea.

Acknowledgements

The authors wish to thank the Korean Central Cancer Registry and National Insurance Data Service for their assistance in disease and mortality data linkage. Moreover, we would also like to thank the study participants who agreed to provide their data.

Conflict of interest: The authors declare that they have no competing interests.

References

1. Emerging Risk Factors Collaboration. Separate and combined associations of body-mass index and abdominal adiposity with cardiovascular disease: collaborative analysis of 58 prospective studies. *The Lancet*. 2011; **377**(9771): 1085-95.
2. 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**(9349): 1903-13.
3. 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**(9602): 1829-39.
4. Mottillo S, Filion KB, Genest J, Joseph L, Poirier P, et al. The metabolic syndrome and cardiovascular risk a systematic review and meta-analysis. *Journal of the American College of Cardiology*. 2010; **56**(14): 1113-32.
5. Sartorius B, Sartorius K, Aldous C, Madiba TE, Stefan C, Noakes T. Carbohydrate intake, obesity, metabolic syndrome and cancer risk? A two-part systematic review and meta-analysis protocol to estimate attributability. *BMJ open*. 2016; **6**(1): e009301.
6. Micucci C, Valli D, Matakchione G, Catalano A. Current perspectives between metabolic syndrome and cancer. *Oncotarget*. 2016; **7**(25): 38959-72.
7. Korean Statistical Information Service (KOSIS). Statistical database. [cited 25th July 2017]; Available from: <http://kosis.kr/eng/>
8. Clarke R, Shipley M, Lewington S, Youngman L, Collins R, Marmot M, et al. Underestimation of risk associations due to regression dilution in long-term follow-up of prospective studies. *American journal of epidemiology*. 1999; **150**(4): 341-53.
9. Sull JW, Kim HJ, Yun JE, Kim G, Park EJ, Kim S, et al. Serum adiponectin is associated with family history of diabetes independently of obesity and insulin resistance in healthy Korean men and women. *Eur J Endocrinol*. 2009; **160**(1): 39-43.
10. Shin HY, Jung KJ, Linton JA, Jee SH. Association between fasting serum glucose levels and incidence of colorectal cancer in Korean men: The Korean Cancer Prevention Study-II. *Metabolism*. 2014; **63**(10): 1250-6.
11. Jee SH, Sull JW, Lee JE, Shin C, Park J, Kimm H, et al. Adiponectin Concentrations: A Genome-wide Association Study. *Am J Hum Genet*. 2010; **87**(4): 545-52.
12. Jung KJ, Won D, Jeon C, Kim S, Il Kim T, Jee SH, et al. A colorectal cancer prediction model using traditional and genetic risk scores in Koreans. *Bmc Genet*. 2015; **16**.
13. Sull JW, Park EJ, Lee M, Jee SH. Effects of SLC2A9 variants on uric acid levels in a Korean population. *Rheumatol Int*. 2013; **33**(1): 19-23.
14. Wen WQ, Kato N, Hwang JY, Guo XY, Tabara Y, Li HX, et al. Genome-wide association studies in East Asians identify new loci for waist-hip ratio and waist circumference. *Sci Rep-Uk*. 2016; **6**.
15. Jee YH, Lee SJ, Jung KJ, Jee SH. Alcohol Intake and Serum Glucose Levels from the Perspective of a Mendelian Randomization Design: The KCPS-II Biobank. *PloS one*. 2016; **11**(9).
16. Jia WH, Zhang B, Matsuo K, Shin A, Xiang YB, Jee SH, et al. Genome-wide association analyses in east Asians identify new susceptibility loci for colorectal

328 cancer. Nature genetics. 2013; **45**(2): 191-6.
329 17. Manolio TA, Collins R. Enhancing the feasibility of large cohort studies.
330 JAMA : the journal of the American Medical Association. 2010; **304**(20): 2290-1.
331
332
333

Figures

Figure 1. Location of the recruiting health facilities in South Korea

Figure 2. Cumulative recruitment into the cohort

Figure 3. Major vascular risk factors by year of birth or age at recruitment

(A) Prevalence of smoking and drinking in men, by birth cohort

(B) Prevalence of smoking and drinking in women, by birth cohort

(C) Prevalence of obesity and diabetes (diagnosed, treated or fasting glucose >7 mmol/L) in men, by age at recruitment

(D) Prevalence of obesity and diabetes (diagnosed, treated or fasting glucose >7 mmol/L) in women, by age at recruitment

(E) Mean systolic and diastolic blood pressure in men and women, by age at recruitment

(F) Mean HDL and non-HDL cholesterol in men and women, by age at recruitment

Table 1. Summary of data collected at baseline (2004-2013) in the Korean Cancer Prevention Study-II Biobank aged 20–84 (N=156 701)

Variable	N	% complete
Social status		
Medical insurance premium	156 701	100
Other socio-economic measures*	38 149	24
Health status		
Medical history (past or family)	156 701	100
Medication history (past or current)	156 701	100
Behavioural characteristics		
Alcohol (type, frequency, and amount)	134 549	86
Smoking (status, amount, duration)	147 020	94
Recreational physical activity	145 961	93
Physiological characteristics		
Anthropometry (weight, height, waist circumference)	156 701	100
Blood pressure	153 673	98
Pulmonary function**	92 587	59
Blood measurements		
Lipid profile (total, HDL, TG, LDL)	156 701	100
Fasting glucose test	156 701	100
Liver function test (GOT, GPT, GGT)	156 701	100
Kidney function (creatinine, BUN)	148 409	95
Albumin, bilirubin	145 310	93
Tumor markers (CEA, CA19-9, CA125)***	90 352	58
Uric acid	143 078	91

HDL: High-density lipoprotein, LDL: Low-density lipoprotein, TG: Triglyceride, GOT: Glutamate oxaloacetate transaminase, GPT: Glutamic pyruvic transaminase, GGT: Gamma-glutamyl transferase, CEA: Carcinoembryonic antigen

* This information was only recorded in 6 of the 18 centres (which together recruited 50 273 of the 156 701 participants). ** Done in 9 of the 18 centres. *** Done in 13 of the 18 centres.

Table 2. Number of deaths, vascular events and cancer incidents according to major ICD-10 categories over 8 years of follow-up in the Korean Cancer Prevention Study-II (to December 2015)

	Men	Women	Total
Number of participants	94 840	61 861	156 701
Number of deaths	869	315	1184
Morbidity (incidence)			
All vascular	4908	2344	7252
Ischemic heart disease	2191	741	2932
Stroke	1357	781	2138
Ischaemic stroke	637	294	931
Subarachnoid haemorrhage	111	66	177
Intracerebral haemorrhage	156	69	225
Other vascular	456	393	849
All cancer	3167	2502	5669
Thyroid*	856	1291	2147
Stomach	523	139	662
Colorectal	371	148	519
Breast	3	429	432
Lung	230	88	318
Prostate	295	-	295
Liver	210	28	238
Kidney	103	21	127
Cervix	-	40	40
Diabetes	4521	1828	6349
End Stage Renal Disease	299	122	421
Dementia	154	95	249
Chronic obstructive pulmonary disease	318	82	400
Tuberculosis	265	133	398

* Thyroid cancer was reported from national cancer centre (NCC) registry, hospital admission records, and death registers (1018) but also detected from ultrasonography (1129) that was available to >100,000 participants as part of their baseline examination.

Table 3. Baseline characteristics of participants recruited into the Korean Cancer Prevention Study-II

	Men (n=94 840)	Women (n=61 861)	Total (n=156 701)
Age, years	42.1 (10)	41.1 (11)	41.7 (11)
Body mass index, kg/m ²	24.4 (3)	22.2 (3)	23.6 (3)
Waist circumference, cms	84.9 (7.8)	74.3 (8.5)	80.7 (9.6)
Diabetes			
Self-reported or treated diabetes	4115 (4.3%)	1493 (2.4%)	5608 (3.6%)
Fasting glucose >7mmol/L	4188 (4.4%)	1265 (2.0%)	5453 (3.5%)
Self-reported or treated diabetes, or fasting glucose >7mmol/L	6116 (6.4%)	2099 (3.4%)	8215 (5.2%)
Blood pressure, mmHg			
Systolic	121 (13)	112 (14)	118 (14)
Diastolic	77 (10)	71 (10)	74 (10)
Cigarette smoking			
Never	26 720 (29%)	51 424 (92%)	78 144 (53%)
Former	24 403 (27%)	2181 (4%)	26 584 (18%)
Current	40 018 (44%)	2274 (4%)	42 292 (29%)
Alcohol consumption			
None	11 452 (14%)	27 389 (54%)	38 841 (29%)
One day per week	22 704 (27%)	12 718 (25%)	35 422 (26%)
More than one day per week	50 096 (59%)	10 153 (20%)	60 286 (45%)
Recreational physical activity			
None	30 456 (34%)	30 281 (54%)	60 737 (42%)
1-3 days per week	50 320 (56%)	20 242 (36%)	70 562 (48%)
≥4 days per week	8 724 (10%)	5 938 (10%)	14 662 (10%)
Blood lipids			
Total cholesterol (mmol/L%)	5.0 (0.9)	4.8 (0.9)	4.9 (0.9)
LDL cholesterol (mmol/L)	3.0 (1.7)	2.8 (0.8)	2.9 (1.4)
HDL cholesterol (mmol/L)	1.3 (0.2)	1.5 (0.3)	1.4 (0.3)

HDL, high-density lipoprotein; LDL, low-density lipoprotein; SD: Standard deviation

Values are mean (SD) or n (%). Percentages are among those with non-missing data (see Table 1 for completeness of data).