








 Data Resource Profile

Data Resource Profile: Understanding the patterns and determinants of health in South Asians—the South Asia Biobank

Peige Song ^{1,2}, **Ananya Gupta**,^{1,3} **Ian Y Goon**,¹ **Mehedi Hasan**,⁴
Sara Mahmood,⁵ **Rajendra Pradeepa**,⁶ **Samreen Siddiqui**,³
Gary S Frost ⁷, **Dian Kusuma** ⁸, **Marisa Miraldo**,^{8,9} **Franco Sassi**,^{8,9}
Nick J Wareham,¹⁰ **Sajjad Ahmed**,¹¹ **Ranjit M. Anjana**,⁶
Soren Brage ¹⁰, **Nita G Forouhi**,¹⁰ **Sujeet Jha**,³
Anuradhani Kasturiratne,¹² **Prasad Katulanda**,¹³ **Khadija I Khawaja**,⁵
Marie Loh,^{1,14} **Malay K Mridha** ⁴, **Ananda R Wickremasinghe**,¹²
Jaspal S Kooner^{15,16} and **John C Chambers**^{1,14*}; on behalf of South Asia Biobank. Remaining authors are listed at the end of the article

¹Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK, ²School of Public Health, Zhejiang University School of Medicine, Hangzhou, China, ³Institute of Endocrinology, Diabetes & Metabolism, Max Super Speciality Hospital (Devki Devi Foundation), New Delhi, India, ⁴Centre for Non-communicable Diseases and Nutrition (CNCDN), BRAC James P Grant of Public Health, BRAC University, Dhaka, Bangladesh, ⁵Department of Endocrinology & Metabolism, Services Institute of Medical Sciences, Services Hospital, Lahore, Pakistan, ⁶Madras Diabetes Research Foundation, Chennai, India, ⁷Faculty of Medicine, Imperial College London, London, UK, ⁸Centre for Health Economics and Policy Innovation, Imperial College Business School, Imperial College London, London, UK, ⁹Department of Economics and Public Policy, Imperial College Business School, Imperial College London, London, UK, ¹⁰MRC Epidemiology Unit, Institute of Metabolic Science, University of Cambridge, Cambridge, UK, ¹¹Punjab Institute of Cardiology, Punjab, Pakistan, ¹²Department of Public Health, Faculty of Medicine, University of Kelaniya, Ragama, Sri Lanka, ¹³Department of Clinical Medicine, Faculty of Medicine, University of Colombo, Colombo, Sri Lanka, ¹⁴Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore, ¹⁵Ealing Hospital, London Northwest University Healthcare NHS Trust, London, UK and ¹⁶National Heart and Lung Institute, Imperial College London, London, UK

*Corresponding author. 172 Medical School, St Mary's Campus, Imperial College London, Norfolk Place, London W2 1PG, UK. E-mail: john.chambers@imperial.ac.uk

Editorial decision 25 January 2021; Accepted 18 February 2021

Key features

- The South Asia Biobank (SAB) is the first comprehensive biobank of South Asian individuals, established to identify the risk factors and their complex interactions underlying the development of type-2 diabetes mellitus (T2DM), cardiovascular disease (CVD) and other chronic diseases in South Asians.
- SAB is a cross-sectional investigation in Bangladesh, India, Pakistan and Sri Lanka, starting in November 2018 and ending in March 2020.
- A total of 52 853 participants took part in SAB, and demographic, lifestyle, clinical, environmental and phenotypic data and biological samples are available.
- Interested research collaborators could refer to the SAB website [<https://www.ghru-southasia.org/>] or contact Professor John C Chambers [john.chambers@imperial.ac.uk].

Data resource basics

Type 2 diabetes mellitus (T2DM) and cardiovascular disease (CVD) are leading and closely interlinked global health challenges.¹ The burdens of T2DM and CVD are especially high in South Asia, one of the most populous and the most densely populated regions of the world.^{2,3} The prevalence of diabetes in South Asia has risen more rapidly than in other large geographical regions,⁴ and it is projected that South Asia will account for 40% of the global CVD burden by 2020.⁵ In addition, T2DM and CVD develop at an earlier age in South Asians than in their European counterparts^{6,7}.

Identification of the primary risk factors for T2DM and CVD is central to the development of effective approaches for the prevention and treatment of chronic diseases such as T2DM and CVD.⁸ However, epidemiological data are currently sparse for South Asia, with evidence on the drivers of T2DM and CVD being predominantly based on cross-sectional studies that recorded a narrow range of exposures and without longitudinal assessments.^{3,5,6,9} The few available prospective studies are largely derived from investigations of South Asians residing in Western countries, and are further limited by small sample size and incomplete phenotypic characterization.³ To better understand the wide range of exposures that contribute to the development of T2DM and CVD in South Asians, a large-scale population-based study that collects information on demographic, lifestyle, clinical, environmental and genomic variables is needed.

To address this important need, we have established a unique cross-sectional population-based study focused on the South Asian population: the South Asia Biobank (SAB). SAB was launched in 2018 as a partnership between collaborating centres in Bangladesh, India, Pakistan, Sri Lanka and the UK.¹⁰ SAB includes rich demographic, lifestyle, clinical, environmental and phenotypic data and biological samples from more than 50 000 South Asian participants. This resource will enable a broad range of epidemiological research, including the development of prevention and treatment approaches, discovery of novel molecular biomarkers, risk stratification algorithms and innovative therapeutic approaches for better prevention of T2DM and CVD in South Asians. The specific initial objectives of SAB are as follows.

- i. Establish a network of non-communicable disease (NCD) surveillance sites in Bangladesh, India, Pakistan and Sri Lanka, using common protocols and platforms, in partnership with regional centres of excellence in South Asia.
- ii. Complete structured health assessments on a representative sample of at least 50 000 South Asians aged 18 years and above, residing at all surveillance sites.
- iii. Use the data to identify the genetic and environmental factors underlying non-communicable diseases in South Asians, and translate the findings into new approaches for maintenance of health and well-being.

The South Asia Biobank is conducted in accordance with the recommendations for physicians involved in research on human subjects, adopted by the 18th World Medical Assembly, Helsinki, 1964, and later revisions. Research approval was obtained from the Imperial College London Research Ethics Committee (reference: 18IC4698) and local institutional review boards in each of the participating countries.

Data collected

SAB is a cross-sectional population-based study that recruited participants from 118 surveillance sites that were centred on local primary community health care units in five study regions: Bangladesh, South India, North India, Pakistan and Sri Lanka. The locations of all surveillance sites are demonstrated in [Figure 1](#). Recruitment started in November 2018 and ended in March 2020 (due to the pandemic of COVID-19).

We recruited men and women of self-reported South Asian ethnicity and aged 18 years and above. We excluded women who were currently pregnant, as well as people who were not permanent residents of the surveillance site (residence for 12 months or more required).

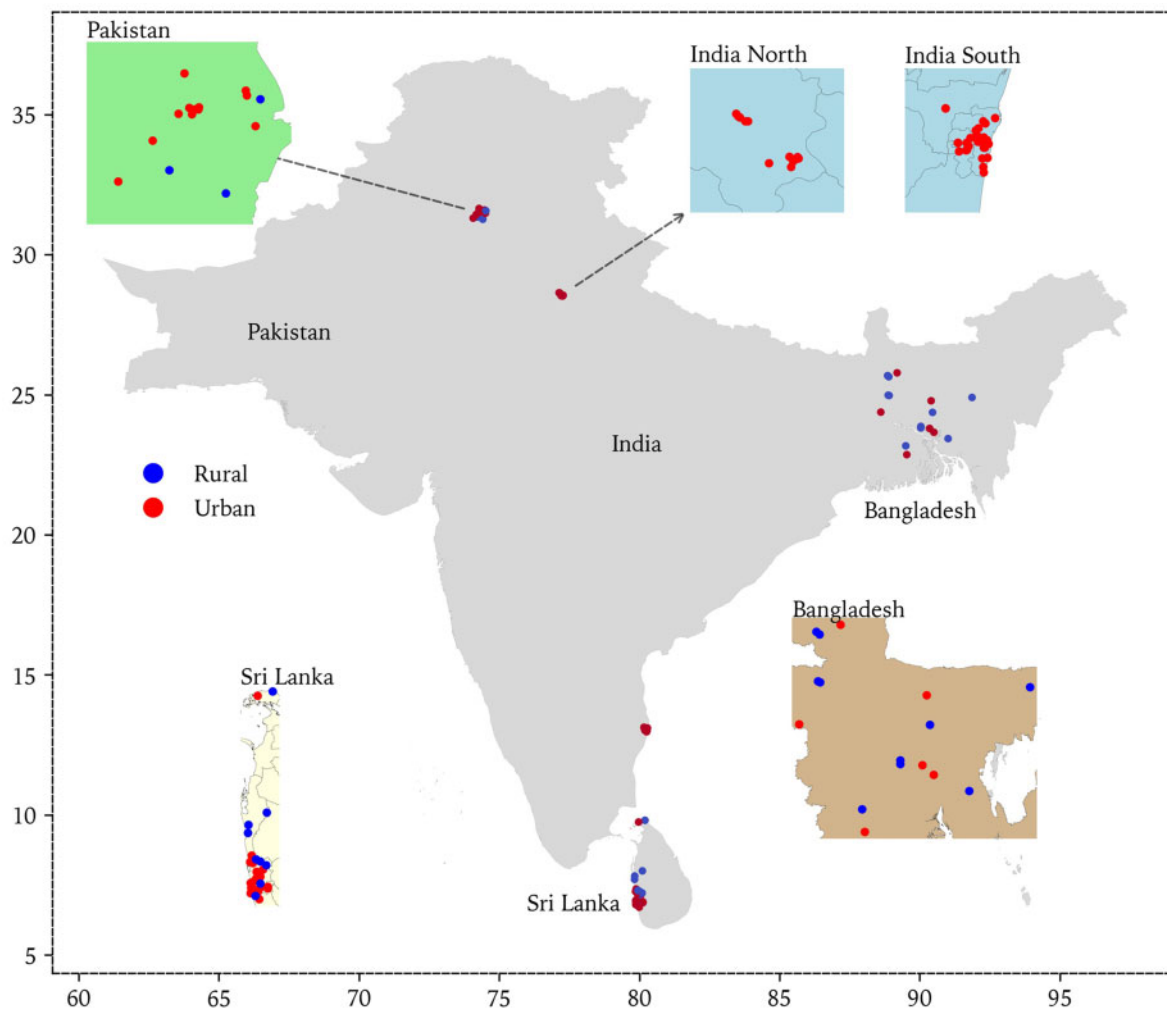


Figure 1. Locations of South Asia Biobank (SAB) surveillance sites

We also excluded people with serious illness expected to reduce life expectancy to less than 12 months, those who planned to leave the surveillance site within the next 12 months and those unable or unwilling to give informed consent.

The surveillance sites at which recruitment occurred in each study region are summarized in [Supplementary Table S1](#) (available as [Supplementary data](#) at *IJE* online). Governmental census data and available household listings were used, together with house-to-house visits by research teams and local primary care workers, to identify (enumerate) the resident population. All individuals in each household who met the inclusion criteria were invited to take part, and their demographic details were obtained in their households. We worked closely with senior community members (e.g. teachers, employers, religious leaders) to support and facilitate engagement in the study. Explanations of the project's purpose were provided in writing and using videos, in relevant South Asian languages, supported by bilingual translators.

By March 2020 we recruited a total of 52 853 subjects: 13 954 from Bangladesh, 8620 from South India, 9469 from North India, 5875 from Pakistan and 14 935 from Sri Lanka. The response rate based on enumerated population in each surveillance site ranged from 17.6% in North India to 72.3 % in Pakistan (see [Supplementary Table S2](#) for more details, available as [Supplementary data](#) at *IJE* online). The demographic structure of the study participants and the comparison with the National Population data in 2015, obtained from the United Nations Population Division, are shown in [Table 1](#).¹¹

Measures

Participants were invited to attend the survey sites between 7 am and 11 am in the fasting state (water only after midnight). Structured assessments of participants were conducted in six complementary domains: (i) Registration and consent; (ii) Health and lifestyle questionnaire; (iii). Physical measurements; (iv) Biological samples (blood and

Table 1 Characteristics of participants in South Asia Biobank (SAB), compared with reported national population distribution (United Nations Population Division)

Demographic characteristic	Bangladesh			India			Pakistan			Sri Lanka		
	National	SAB (n = 13 954)	P-value	National	SAB (n = 18 089)	P-value	National	SAB (n = 5875)	P-value	National	SAB (n = 14935)	P-value
Age			<0.001			<0.001			<0.001			<0.001
<30 years	30.28	15.71		28.53	13.07		34.32	15.74		21.17	10.56	
30-50 years	45.24	50.61		42.93	47.95		41.66	51.56		40.89	39.17	
50-70 years	19.08	29.25		23.09	33.37		19.04	30.88		29.98	40.74	
≥70 years	5.40	4.44		5.45	5.62		4.98	1.82		7.96	9.53	
Sex			<0.001			<0.001			<0.001			<0.001
Male	50.45	55.43		51.57	61.53		51.11	67.09		47.18	69.07	
Female	49.55	44.57		48.43	38.47		48.89	32.91		52.82	30.93	

Data are in percentages. The age category of '<30 years' refers to '20-29 years' in the United Nations Population Division and '18-29 years' in SAB; chi-square goodness of fit testing was used to determine whether the age and sex proportions in SAB were similar to those in the general population.

spot urine); (v) Physical activity monitoring; and (vi) 24-h dietary recall. Procedures and training were standardized between countries and surveillance sites.

Registration and consent

Written, informed consent was obtained from all participants for data collection and inclusion in the research. Informed consent included permission for the data and samples collected to be used for chronic disease research, including data sharing with national and international bodies concerned with prevention and control of T2DM and CVD and for molecular epidemiological research. Consent was facilitated using videos (available in major South Asian languages). A unique study identity number (ID) was allocated to each participant.

Questionnaire

An interviewer-administered health and lifestyle questionnaire was used to collect information on behavioural risk factors (smoking, alcohol use, physical activity and consumption of fruits/vegetables), personal and family medical history, medications and socioeconomic status. The questionnaire was founded on the extended World Health Organization (WHO) STEPwise approach to surveillance questionnaire that is widely used in global disease surveillance and which was adapted for use in the South Asia context, through incorporating additional questions.¹²

Physical measurements

These included: (i) anthropometry (height, weight, waist and hip circumference and bio-impedance for body fat composition); (ii) blood pressure by digital device; (iii) cardiac evaluation by 12-lead electrocardiogram to identify arrhythmia, left ventricular hypertrophy and previous myocardial infarction; (iv) retinal photography for

assessment of retinal disease, including hypertensive and diabetic retinopathy; and (v) respiratory evaluation by spirometry to assess for smoking/environment-related lung injury.

Biological samples

Using venesection by trained phlebotomists, 25 ml venous blood was collected and then distributed into ethylenediaminetetraacetic acid (EDTA), serum and citrate vacutainer tubes, and into tubes designed for RNA preservation (Tempus tube). Fasting glucose and cholesterol were measured by point of care tests. An oral glucose tolerance test was carried out in a subset of participants, enabling validation of diabetes classification. A spot urine sample (6 ml in three aliquots) was also collected for analysis of albuminuria and other biomarkers. Aliquots of whole blood, buffy coat, serum, EDTA plasma, citrate plasma and urine (Supplementary Table S3, available as Supplementary data at *IJE* online) were stored at -80°C for future molecular epidemiological research (including genomics), to investigate the mechanisms underpinning the development of T2DM, CVD and other complex diseases that are of importance to South Asians (including but not limited to: obesity, cancer, dementia, chronic obstructive pulmonary disease, chronic kidney disease).

Physical activity

This was also objectively quantified in 100-Hz resolution using a wrist-worn triaxial accelerometer, worn on participants' non-dominant wrist for 7 days. This device is small, light-weight, wrist-watch-shaped, battery-powered and uses triaxial acceleration in gravitational units to infer participant movement. It has been used recently to measure physical activity patterns amongst 100 000 people in the UK Biobank study.¹³

Dietary intake

Consumption was recorded by interviewer-administered computerized 24-h dietary recall based on the multiple pass method using the Intake24 system [<https://intake24.org/>]. The system was specifically adapted for the South Asian context through incorporating extensive additional foods, drinks and dishes, and portion-size photographs relevant to the study settings. Adaptation was informed by research nutritionists and dieticians from each study centre and by the results of previous dietary surveys in the study locations. The implementation of this tool could enable the description of food and nutrient intakes, evaluation of intakes in comparison with guidelines and investigation of the link between diet and health endpoints.

All study participants received a report summarizing the clinically relevant results of their health assessment, together with an explanatory booklet and a link to access an explanatory video. Participants identified with significant health conditions (e.g. T2DM, hypertension) had the opportunity to discuss the results with the study team, and to be referred to an appropriate health care facility for further assessment, counselling or treatment.

Environmental mapping

In each surveillance site, an environmental mapping exercise was carried out. The aim was to characterize the built environment in terms of retailers and advertisements for food and tobacco and physical activity facilities. The methodologies were adapted from food modules conducted by: the International Network for Food and Obesity/NCDs Research, Monitoring and Action Support; the Maryland Food Systems Map conducted by the Johns Hopkins Center for a Livable Future; and the World Health Organization Framework Convention on Tobacco Control.^{14–16} In addition to geolocations, the main variables included food (e.g. fruit, vegetables, confectionery), drinks (e.g. soft drinks, sugar-free drinks) and tobacco products (e.g. cigarette, beedi) being sold or advertised. Data collection used KoboToolBox for Android [<https://www.kobotoolbox.org>] and covered each surveillance site with a 500-m buffer beyond the site boundary.

Identification of outcomes

The primary outcomes were T2DM and cardiovascular disease. The secondary endpoints included respiratory and chronic kidney diseases, or cancer.

Quality control and data management

The surveillance teams, comprising research assistants, laboratory technicians, physicians and coordinators, were

trained to follow standardized protocols (Supplementary Table S4, available as Supplementary data at *IJE* online). Their training modules included interviewing techniques, ethics and specific instructions for data variables (demographic, socioeconomic, food security, behavioural risk factors, medication and lifestyle practices, physical measurement and collection of biological samples).

Revalidation of the research teams in study procedures was done at regular intervals during the study to ensure high-quality data collection that was harmonized across surveillance sites. Standardized operating procedures were established for all data collection procedures. Questionnaires were translated into the local languages local to the communities, and back-translated. Equipment used for physical and biological measurements is listed in Supplementary Table S5, available as Supplementary data at *IJE* online, and was regularly calibrated using appropriate controls/standards.

The data management teams reviewed the data collected routinely for completeness and data quality, including using custom computer scripts to assess for biases in data entry, logical inconsistencies, internal correlations, digit preference and measurement drift or bias between machines and observers. Quality control reports were circulated at weekly intervals between the study investigators, to drive continuous evaluation and improvement in study processes. A random subset comprising up to 2% of the study participants, and/or a subset of biological samples, was reassessed to provide additional quality control information. Data collection methods used were ‘field-friendly’, culturally acceptable and minimally invasive in order to reduce participant attrition and improve logistical feasibility.

Personal and clinical data were separated by pseudonymization to enhance data security. All data were encrypted during transmission and stored securely both locally and in a cloud-based infrastructure. Data and all relevant documents will be stored for a minimum of 10 years. Samples collected were split and stored in both South Asia and the UK to ensure long-term (>20 years) sample integrity and preservation. Some laboratory assays on stored samples were done in South Asia and the majority of assays were carried out in the UK or other countries with relevant technologies in the future.

Data resource use

Data collected in this cross-sectional investigation could be used to assess the epidemiology of T2DM, CVD and other chronic diseases in South Asia. The exploration of possible risk factors for T2DM and CVD could provide a scientific basis for evidence-based public health policy making and

interventions. Upon request, the rich resources of SAB are available to researchers from all over the world.

Based on the definition of obesity by WHO, the prevalence of obesity (body mass index ≥ 30 kg/m²) is 6.6% in Bangladesh, 19.7% in India, 33.9% in Pakistan and 15.7% in Sri Lanka. Similarly the prevalence of diabetes, defined as a fasting glucose level >126 mg/dL or a physician diagnosis or current antidiabetic medication, is 11.5%, 27.7%, 25.3% and 24.8%, respectively; and the prevalence of hypertension, defined as a systolic blood pressure ≥ 140 mmHg or a diastolic blood pressure ≥ 90 mmHg or a physician diagnosis or current antihypertensive medication, is 26.7%, 36.9%, 44.5%, 35.0% in Bangladesh, India, Pakistan and Sri Lanka, respectively.

Strengths and weaknesses

SAB is designed to identify the risk factors and their complex interactions underlying the development of T2DM, CVD and other chronic diseases in South Asians. With intensive data collection, SAB provides representative population samples in four South Asian countries. SAB is the first comprehensive biobank of South Asian individuals. Its large sample size, broad geographical reach and wide range of data collected, including biosamples, make SAB a powerful tool for epidemiological and translational research in South Asian populations. The standardized procedures and rigorous quality control of data collection ensure comparability of study results between and within the partner countries. Further, although random sampling approaches were used in selecting participants, we cannot exclude 'healthy volunteer' effects, a common phenomenon in epidemiological research. In addition, advanced phenotyping by imaging (e.g. MRI, DXA or ultrasound) was not feasible across the range of sites studied.

Data resource access

Reports and major results of SAB will be released regularly on the SAB website [<https://www.ghru-southasia.org/>]. Any enquiries regarding SAB should be directed to Professor John C Chambers [john.chambers@imperial.ac.uk]. Subject to data privacy requirements and the permissions included in the consent form, individual-level data and samples are available for use to approved investigators.

Supplementary data

Supplementary data are available at *IJE* online.

Acknowledgements

All authors thank all the team members and all participants in the South Asia Biobank. Remaining authors from the South Asia Biobank are: Polly Page,¹ Wnurinham Silva,² Garudam R Aarthi,³ Saira Afzal,⁴ Sophie E Day,^{2,5} Bridget A Holmes,¹ Rajan Kamalesh,³ Elisa Pineda,⁶ Fred Hersch,⁷ Baldeesh K Rai,² Malabika Sarker,⁸ and Jonathan Valabhji.^{9–11}

¹MRC Epidemiology Unit, Institute of Metabolic Science, University of Cambridge, Cambridge, UK; ²Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK; ³Madras Diabetes Research Foundation, Chennai, India; ⁴King Edward Medical University, Punjab, Pakistan; ⁵Goldsmiths University, London, UK; ⁶Centre for Health Economics and Policy Innovation, Imperial College Business School, Imperial College London, London, UK; ⁷Sydney Medical School, University of Sydney, Sydney, NSW, Australia; ⁸BRAC James P Grant of Public Health, BRAC University, Dhaka, Bangladesh; ⁹Faculty of Medicine, Imperial College London, London, UK; ¹⁰Department of Diabetes and Endocrinology, Imperial College Healthcare NHS Trust, London, UK; ¹¹NHS England & Improvement, London, UK

Funding

The South Asia Biobank is supported by the UK National Institute for Health Research (award number 16/136/68), and by the Wellcome Trust (award number 212945/Z/18/Z). The views expressed are those of the author(s) and not necessarily those of the National Institute for Health Research, the Wellcome Trust or the Department of Health. J.C. is also supported by the Singapore Ministry of Health's National Medical Research Council under its Singapore Translational Research Investigator (STaR) Award (NMRC/STaR/0028/2017). SB, NGF, BH, PP and NJW acknowledge funding from the Medical Research Council Epidemiology Unit (MC_UU_00006/1, MC_UU_00006/3 and MC_UU_00006/4) and from NIHR Biomedical Research Centre Cambridge: Nutrition, Diet, and Lifestyle Research Theme (IS-BRC-1215-20014).

Conflict of interest

None declared.

References

1. Roth GA, Johnson C, Abajobir A *et al*. Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. *J Am Coll Cardiol* 2017;70:1–25.
2. Misra A, Tandon N, Ebrahim S *et al*. Diabetes, cardiovascular disease, and chronic kidney disease in South Asia: current status and future directions. *BMJ* 2017;357:j1420.
3. Dans A, Ng N, Varghese C, Tai ES, Firestone R, Bonita R. The rise of chronic non-communicable diseases in southeast Asia: time for action. *Lancet* 2011;377:680–89.
4. Ghaffar A, Reddy KS, Singhi M. Burden of non-communicable diseases in South Asia. *BMJ* 2004;328:807–10.
5. Gholap N, Davies M, Patel K, Sattar N, Khunti K. Type 2 diabetes and cardiovascular disease in South Asians. *Prim Care Diabetes* 2011;5:45–56.

6. Forouhi N, Merrick D, Goyder E *et al.* Diabetes prevalence in England, 2001—estimates from an epidemiological model. *Diabet Med* 2006;**23**:189–97.
7. Forouhi N, Sattar N, Tillin T, McKeigue P, Chaturvedi N. Do known risk factors explain the higher coronary heart disease mortality in South Asian compared with European men? Prospective follow-up of the Southall and Brent studies, UK. *Diabetologia* 2006;**49**:2580–88.
8. Eckel RH, Kahn R, Robertson RM, Rizza RA. Preventing cardiovascular disease and diabetes: a call to action from the American Diabetes Association and the American Heart Association. *Circulation* 2006;**113**:2943–46.
9. World Health Organization. *Noncommunicable Diseases in the South-East Asia Region, 2011: Situation and Response*. Geneva: World Health Organization, 2012.
10. Sudlow C, Gallacher J, Allen N *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015;**12**: e1001779.
11. United Nations. Department of Economic and Social Affairs, Population Division. *World Population Prospects 2019*. New York, NY: United Nations, 2019.
12. World Health Organization. *WHO STEPS Surveillance Manual: the WHO Stepwise Approach to Chronic Disease Risk Factor Surveillance*. Geneva: World Health Organization, 2005.
13. Doherty A, Jackson D, Hammerla N *et al.* Large scale population assessment of physical activity using wrist worn accelerometers: the UK biobank study. *PLoS One* 2017;**12**:e0169649.
14. INFORMAS. *International Network for Food and Obesity/ Non-communicable Disease Research, Monitoring and Action Support*. <http://www.informas.org/> (20 July 2019, date last accessed).
15. Misiaszek C, Buzogany S, Freishtat H, *Baltimore City's Food Environment: 2018 Report*. Baltimore, MD: Johns Hopkins Center for a Livable Future, 2018.
16. World Health Organization. *Report on the Global Tobacco Epidemic*. The MPOWER package. Geneva: World Health Organization, 2008.