BMJ Open

Kanagawa Investigation of the Total Check-up Data from the National database (KITCHEN): protocol for data-driven population-based repeated cross-sectional and 6-year cohort studies

Kei Nakajima,[1,2] Taizo Iwane,[1] Ryoko Higuchi,[1] Michi Shibata,[1,3] Kento Takada,[1] Jun Uda,[4] Mami Anan,[1] Michiko Sugiyama,[1] Teiji Nakamura[1]

For numbered affiliations see end of article.

Correspondence to
Professor Kei Nakajima;
nakajima-rsh@kuhs.ac.jp

ABSTRACT

Introduction The unmitigated incidence of cardiometabolic diseases, such as type 2 diabetes and metabolic syndrome, has gained attention in Japan. 'Big data' can be useful to clarify conflicting observations obtained from studies with small samples and about rare conditions that are often neglected. We epidemiologically address these issues using data from health check-ups conducted in Kanagawa Prefecture, the prefecture with the second largest population in Japan, in the Kanagawa Investigation of the Total Check-up Data from the National Database (KITCHEN).

Methods and analysis This research consists of a series of population-based cross-sectional studies repeated from 2008–2014 and 6-year cohort studies. Since 2017, we have reviewed the data of people living in Kanagawa Prefecture who underwent a health check-up mainly for general health and the prevention of metabolic syndrome. The sample size ranges from 1.2 million to 1.8 million people in the cross-sectional studies and from 370 000 to 590 000 people in the cohort studies. These are people aged 40–74 years, whose clinical parameters were measured and who responded individually to a questionnaire. We investigate potential associations and causalities of various aetiologies, including diabetes and metabolic syndrome, using clinical data and lifestyle information. With multidisciplinary analysis, including data-driven analysis, we expect to obtain a wide range of novel findings, to confirm indeterminate previous findings, especially in terms of cardiometabolic disease, and to provide new perspectives for human health promotion and disease prevention.

Ethics and dissemination Ethical approval was received from the Ethics Committee of Kanagawa University of Human Services (10-43). The protocol was approved in December 2016 by the Japanese Ministry of Health, Labour and Welfare (No. 121). The study results will be disseminated through open platforms including journal articles, relevant conferences and seminar presentations.

Strengths and limitations of this study

► The number of subjects in the sample is so large that more precise results concerning the means of parameters can be obtained, even when the subjects are classified into multiple categories, including sex, age group, smoking status and certain morbidities such as obesity or diabetes.

► It may be possible to use big data to evaluate minor or rare conditions or aetiologies that are commonly overlooked, neglected or unfeasible to analyse in clinical studies, particularly those with small samples.

► It may also be possible to conduct data-driven and hypothesis-generating studies and then detect latent relationships among measures available in the data.

► Identical measurements and assessments of anthropometric indices, blood pressure, blood biochemistry and urinalysis are performed across multiple years in the prefecture with the second largest population in Japan.

► The variations in parameters are restricted, and parameters for specific diseases are not included, because the check-ups are conducted for general health and the prevention of common diseases, especially lifestyle-related diseases such as type 2 diabetes and metabolic syndrome.

INTRODUCTION

Over the past several decades, the incidence of cardiometabolic diseases such as type 2 diabetes and metabolic syndrome (MetS) has not been reduced and has gained attention in Asia, including Japan,[1 2] which has also experienced an unprecedented acceleration of societal ageing.[3 4] These issues may also be problematic in Kanagawa Prefecture (figure 1), an eastern district of Japan, located near Tokyo. The primary causes of these diseases include unfavourable lifestyles (eg, smoking, heavy alcohol consumption, insufficient sleep and infrequent exercise) and excess body weight (overweight and obesity) because of overeating, along with individuals'
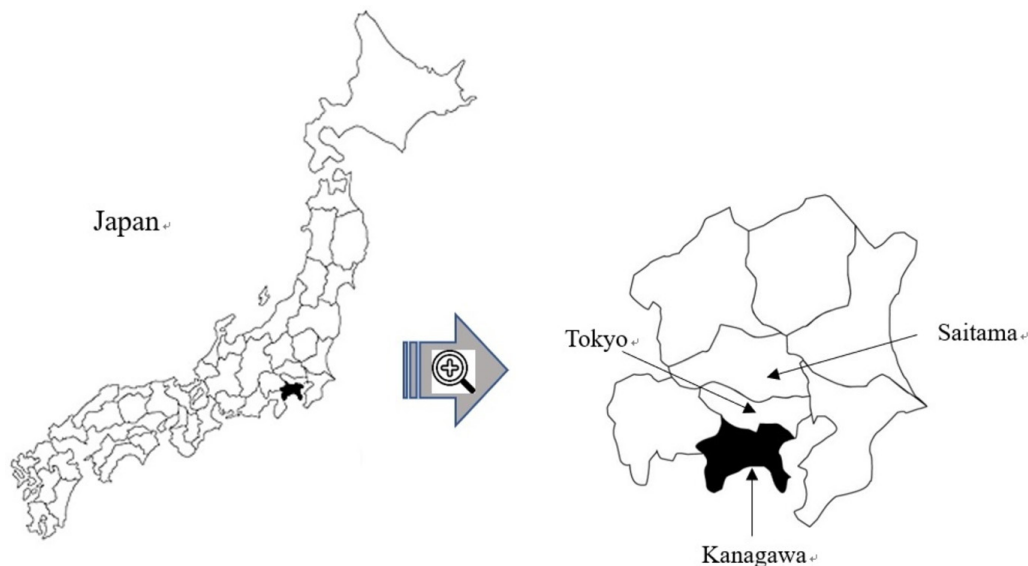
Figure 1 Location of kanagawa prefecture.

genetic and epigenetic backgrounds. However, for the last decade, malnutrition (eg, low body weight) has been shown to be prevalent among young women[5–7] and the elderly[8–10] in Japan, which may contribute to the increased rates of sarcopenia and frailty in the country. Combined with prolonged longevity, cardiometabolic diseases with age-related causes create a long-term burden that leads to direct (ie, measurements and therapies) and indirect (eg, nursing, care and welfare) medical costs nationwide,[11 12] particularly when severe complications such as organ failure (eg, heart, liver and renal failure) develops over the life course.

In 2008, a special health check-up was initiated, primarily for the prevention of MetS, by the Ministry of Health, Labour and Welfare (MHLW) in Japan.[13] Since that time, all people living in Japan aged 40–74 years are supposed to undergo a yearly health check-up. The data from these check-ups have continuously accumulated, creating a very large database. Such 'big data' are likely to be useful in clarifying indistinct or conflicting results obtained from clinical studies with small sample sizes,[14 15] confirming established results and advancing them by elucidating plausible mechanisms and clinical relevance and enabling a precise understanding of the current status of public health and the contributing to it. Additionally, 'big data' of this kind enable us to investigate minor or rare conditions and aetiologies,[14] such as extremely low and high body weight, abnormal (low and high) clinical measurements, and the low or high prevalence of unfavourable habits and lifestyles, where aetiologies are hardly understood, especially when such extreme conditions are combined in complicated ways, primarily because of inadequate numbers of observations and corresponding cases. Extreme conditions can feasibly be reproduced in animal or cellular studies by means of intentional manipulation of conditions including through the use of transgenic and knockout technologies. These non-human laboratory

studies can provide profound insight into the aetiology of human diseases.[16 17] Clearly, such extreme conditions are mostly unfeasible in studies involving humans. However, in an epidemiological study with a database equivalent to 'big data', it might be feasible to reproduce such extreme conditions in certain categories for limited conditions.

Although cardiometabolic diseases such as type 2 diabetes and chronic organ failures such as chronic kidney disease have been increasing along with the prolonged longevity in Japan,[3 4] the underlying associations with clinical parameters and their mechanisms have not been fully elucidated or confirmed, particularly in epidemiological studies using the 'big data' from the health check-ups described above. These data include more than one million observations per year in most prefectures in Japan. To date, no investigation of this type has been performed, especially on the prefecture scale in Japan.

To this end, we investigated current cardiometabolic disease and health status as clearly as possible, as well as the relationship of cardiometabolic diseases, including but not limited to type 2 diabetes and MetS, and age-related aetiologies. We focused especially on the thorough, end-to-end analysis of the variables of interest, using digitally recorded accumulated data in an extremely large epidemiological study of Kanagawa Prefecture, the second most populated prefecture in Japan, with approximately 9 million inhabitants, second only to Tokyo (approximately 13.7 million inhabitants), as of October 2017. Taking this approach, our study may be characterised as a data-driven and hypothesis-generating study, with the nature of 'big data' research, rather than a hypothesis-testing, traditional epidemiological study.[18 19] Consequently, the concrete objectives and contents of individual studies are difficult to determine before it becomes clear what kinds and amounts of data are available, which will substantially influence the design and analysis methods

of each study. Although big data is often analysed with various algorithms including machine learning,[18] in this study, we analysed the data using traditional the epidemiological methods described in the next section, rather than machine learning.

METHODS AND ANALYSIS
Design
In 2013, the MHLW began to offer accumulated data consisting of information on patient prescriptions and health check-ups for use by Japanese institutions including universities, hospitals and research centres. These data are recorded digitally and are provided in a third-party manner, according to the concept of the 'provision of medical-related data to a third party' to improve the quality of medical services and to support academic research in Japan.[20] To date, 178 applications from various institutions in Japan have been accepted in this manner (as of 30 March 2018).

Our project was a composite multidisciplinary study aimed at elucidating the factors associated with cardiometabolic diseases and eventually contributing to the amelioration and advancement of social health and welfare. After the study protocol was approved by the ethics committee of Kanagawa University of Human Services (10-43), we applied to the MHLW's data provision system in October 2016, through Teiji Nakamura, the president of Kanagawa University of Human Services, as a representative. The protocol of our study was approved in December 2016 by the MHLW (No. 121), after a peer review by an expert council.

Before we received the database from the MHLW, identifying individual-level information (names and postal codes) was completely transformed into randomised non-distinguishing anonymous numbers and characters, which prevents the restoration of this information by any means. There are two types of unique identifying variables available for each subject in the cross-sectional database collected from 2008 to 2014: ID 1 is determined based on the subject's insurance number, sex and birth date, and ID 2 is determined by the subject's name, sex and birthday. Both variables consist of anonymous numbers and characters created by the MHLW using a hash function.[21] For individual subjects, these variables are unchanged in principle, except when there are changes in the variables' constituting parts.

To further protect against the identification of specific individuals, age was categorised into 5-year age groups (40–44, 45–49, 50–54, 55–59, 60–64, 65–69 and 70–74 years), so that the individual's precise age at the time of data collection is unknown in our study.

Our study is part of the MHLW's nationwide programme of providing medical-related data to third parties,[20] and informed consent for the use of these data has not been obtained from each subject. We have opened the protocol of our study to the public on our university homepage, which was updated in October 2017,[22] in line with the
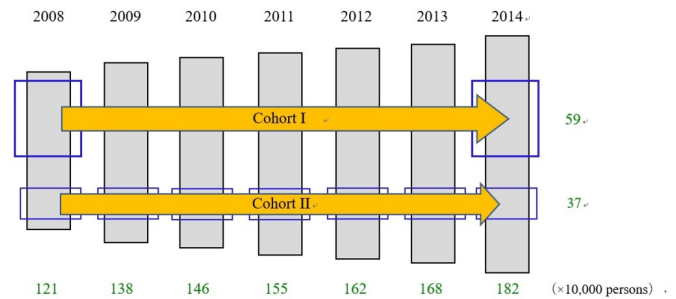

Figure 2 Structure of the cross-sectional and cohort studies. They grey rectangles represent the each year's cross-sectional study. Cohort study I consists of the cross-sectional studies of 2008 and 2014, and cohort study II includes all years from 2008 to 2014. The numbers highlighted in green represent the sample size of each dataset.

'Ethical Guidelines for Medical and Health Research Involving Human Subjects'[23] in Japan (updated by the MHLW and the Ministry of Education, Culture, Sports, Science and Technology in May 2017). We received the digitally recorded non-distinguishing anonymous data from the MHLW in August 2017.

Our analysis of the data was conducted in a location with restricted access and tight security regarding datasets at Kanagawa University of Human Services. Repeated cross-sectional studies will be conducted using check-up data from 2008 to 2014. Additionally, a historical cohort study will be conducted, using the 2008 data as a baseline and the 2014 data to assess final outcomes (figure 2). During this period, the number of subjects undergoing a health check-up has increased each year in Kanagawa in parallel with the nationwide trend. Nationally, almost 50% of the population attended a health check-up in 2014, probably because of the political encouragement for these check-ups (figure 3), although the MHLW's overall expected target rate is 70%.[24]

People who did not undergo a check-up might have been under treatment for moderate to severe disease or hospitalised at the relevant time points. Health-minded people in Japan were likely to voluntarily undergo an expensive health check-up called the '*Ningen Dock*' (detailed and comprehensive health check-up). Other people who did
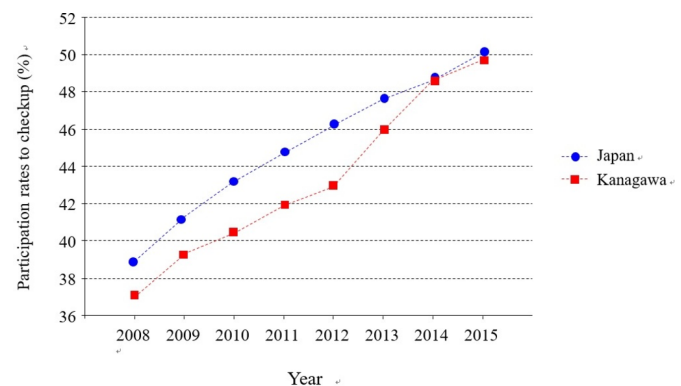
Figure 3 Check-up participation rates (%).

not undergo a check-up might have missed the opportunity to have a check-up because of business obligations or other reasons, including family reasons or moving.

The name of the full study is the Kanagawa Investigation of the Total Checkup data from the National (KITCHEN) database. Each subsequent KITCHEN publication will be numbered sequentially from 1.

Subjects and measurements

People aged 40–74 years and living in Kanagawa Prefecture at the time of the data collection were enrolled in a series of studies. Those residing in medical institutions including hospitals and nursing homes were not included. All subjects are thought to be active to the extent of coming to the place where the check-up was performed. However, some of the subjects have diseases such as hypertension, diabetes or dyslipidaemia, and some have a history of morbidities such as heart disease or stroke. All of these conditions are digitally recorded as answers to a questionnaire. Specific exclusion and inclusion criteria will be determined for each study in the future. The sample sizes range from 1.2 million to 1.8 million people in the cross-sectional studies and from 370 000 to 590 000 people in the cohort studies (figure 2). Cohort study I uses data from 590 000 people who attended check-ups in 2008 and 2014. Cohort study II is based on data from 370 000 people who attended a check-up every year from 2008 to 2014. The two types of identifying variables (ID 1 and ID 2) described above are used to link data on individual subjects throughout the cohort study. When both of these variables simultaneously changed for a subject from 2008 to 2014, it was not possible to follow these individuals through time, resulting in the exclusion of these subjects from the cohort study. To date, such an event has been reported to occur at a rate of approximately 0.8% per year.[21]

Patient and public involvement

Patients are not involved in this study.

All of the parameters measured in this study are listed in table 1. To evaluate subject's age as a numeric value, we transformed age group (40–44, 45–49, 50–54, 55–59, 60–64, 65–69 and 70–74 years) into substituted age (s-age), corresponding to the median for each age group (42, 47, 52, 57, 62, 67 and 72 years). Body weight and height were objectively measured by trained institutional staff members and were recorded to one decimal place (kg and cm). Body mass index (BMI) was calculated as weight (in kg) divided by height (in m^2). In most cases (approximately 99.9%), waist circumference (WC) was measured objectively at the navel level by a medical staff member and recorded to one decimal place. Biochemical measurements were performed using standard methods and automated machines. Dipstick urine analysis for proteinuria and glycosuria was assessed visually or with ordinary automated machines. Several different methods were used for the included biochemical parameter (table 1). The measurement of blood pressure and

blood/urine biomarkers was regularly standardised using both internal standards with available traceability and external standards by third parties, including the Japanese Association of Medical Technologists, even when the measurements were outsourced.[25]

In principle, most people underwent a check-up after overnight fasting. However, some of the check-ups were conducted in a non-fasting condition because of, for example, shift work or family reasons. Therefore, all subjects were asked for the time (in hours) from their last meal to the time of the check-up, which was recorded as at least 10 hours or less than 10 hours. Those completing the check-up less than 10 hours after their last meal will be distinguished from others in certain substudies, for instance, when examining diabetes or dyslipidaemia.

The Japanese diagnostic criteria for MetS were published in 2005.[26] Unlike other criteria such as that of the Adult Treatment Panel III (ATP-III) and the International Diabetes Federation (IDF),[27 28] the Japanese MetS criteria include abdominal obesity as an essential condition (WC ≥85 cm for men and ≥90 cm for women), in addition to two or more of the following three components: (1) dyslipidaemia (triglycerides ≥150 mg/dL and/or high-density lipoprotein cholesterol <40 mg/dL, or pharmacotherapy for dyslipidaemia); (2) hypertension (systolic blood pressure ≥130 mm Hg and/or diastolic blood pressure ≥85 mm Hg, or pharmacotherapy for hypertension); and (3) hyperglycaemia (fasting plasma glucose (FPG) ≥110 mg/dL or pharmacotherapy for diabetes). In the practice of the health check-ups, hyperglycaemia is defined as having elevated FPG (≥110 mg/dL) and/or HbA1c (National Glycohemoglobin Standardization Program [NGSP]) ≥6.0% or pharmacotherapy for diabetes. Furthermore, pre-MetS is defined as abdominal obesity plus one of the three components listed above.[24] In substudies concerning MetS, MetS will also be determined using other international criteria, such as that of the ATP-III or the IDF to allow for comparison with the same criteria with other Asian countries as well as with Western countries. HbA1c (Japan Diabetes Society [JDS]) was converted to HbA1c (NGSP) units using the officially certified formula: HbA1c (NGSP) (%)=1.02 × JDS (%)+0.25%.[29] In 2008, almost all subjects had either FPG or HbA1c measured (99.7%).

If the data on serum creatinine eventually become available, the estimated glomerular filtration rate (eGFR) will be calculated using the s-age above and the following equation[30]: eGFR (ml/min/1.73 m^2)=194 × serum Cr$^{-1.094}$ × s-age$^{-0.287}$ (if female)×0.739, where Cr denotes serum creatinine concentration (mg/dL).

Hypertensive retinopathy has been shown to be associated with cardiovascular events and mortality.[31 32] Hypertensive retinopathy assessments using the Keith-Wagener and Scheie classifications are available in the study, although a very small percentage of individuals (around 1.3%) completed the hypertensive retinopathy examination.

Table 1 Clinical characteristics and methods for measurements

Parameters		Data of 2008 (total n=1 209 118)		Methods and remarks
Substituted age (years old)*		55.0±9.8 in total. 53.9±9.5 for men. 56.6±10.1 for women.		Actual age is unknown but classified into every 5 years.
Sex, n (%)		Men: 695 055 (57.5). Women: 514 063 (42.5).		
Anthropometric parameters				
		Means±SD	Ratios of methods or categories (%)	
BMI (kg/m^2)		23.0±3.3 in total. 23.7±3.1 for men. 22.0±3.3 for women.	–	Weight (in kg) divided by height (in m^2).
Waist circumference (cm) (1 200 959)[†]	Men (690 133)[†]	84.5±8.4	–	Objectively measured.
	Women (510 826)[†]	79.3±9.6	–	
MetS (%) (1 201 807)[†]		MetS	13.4	The diagnosis is determined by Japanese criteria.[19]
		Pre-MetS	13.1	
		Non-MetS	72.2	Including non-pre-MetS.
		Unknown	1.4	Due to incomplete data.
Health guidance level (1 200 272)[†]		Intensive health guidance	11.1	Abdominal obesity‡ + risk factors§ ≥2 or risk factors§ ≥3 + BMI ≥25 kg/m^2 (without abdominal obesity).
		Motivational health guidance	9.6	Abdominal obesity‡ + 1 risk factor§ or one or two risk factors§ + BMI ≥25 kg/m^2 (without abdominal obesity).
		Not applicable	77.9	
		Unknown	1.5	
Systolic blood pressure (mm Hg)		125±17	68.9	Once measurement.
		125±18	24.2	First time measurement among twice measurements.
		131±19	6.9	Second time measurement among twice measurements.
Diastolic blood pressure (mm Hg)		77±11	67.5	Once measurement.
		77±12	24.2	First time measurement among twice measurements.
		81±13	8.2	Second time measurement among twice measurements.
Serum parameters				
Aspartate aminotransferase (U/L) (1 208 753)[†]		24±12	84.9	Ultraviolet spectrophotometric determination (Japan Society of Clinical Chemistry (JSCC reference method).
		23±11	15.1	Others.
Alanine aminotransferase (U/L) (1 208 455)[b]		23±17	84.9	Ultraviolet spectrophotometric determination (JSCC reference method).
		24±18	15.1	Others.
Gamma-glutamyl transferase (U/L) (1 208 074)[†]		36±31	82.6	Ultraviolet spectrophotometric determination (JSCC reference method).
		36±31	17.4	Others.
Triglyceride (mg/dL) (1 208 775)[†]		96 (68–141)	81.5	Ultraviolet and visible spectrophotometric determination (enzyme colorimetric/glycerol elimination methods).
		97 (68–146)	3.2	Ultraviolet spectrophotometric determination (enzyme colorimetric/glycerol elimination method).
		96 (66–145)	15.3	Other.
High-density lipoprotein cholesterol (mg/dL) (1 208 872)[†]		64±17	80.6	Ultraviolet and visible spectrophotometric determination (direct methods [non-precipitation method]).
		61±16	3.7	Ultraviolet spectrophotometric determination (direct methods [non-precipitation method]).
		63±17	15.7	Other.
Low-density lipoprotein cholesterol (mg/dL) (1195,947)[†]		127±31	80.8	Ultraviolet and visible spectrophotometric determination (direct methods [non-precipitation method]).
		125±31	3.4	Ultraviolet spectrophotometric determination (direct methods [non-precipitation method]).
		125±31	15.9	Other.
Uric acid (mg/dL)¶		NA	NA	NA

Continued

Table 1   Continued

Parameters	Data of 2008 (total n=1 209 118)		Methods and remarks
Creatinine (mg/dL)¶	NA	NA	NA
eGFR (mL/min/1.73 m²)¶	NA	NA	NA
Fasting plasma glucose (mg/dL)** (993 458)‡	98±19	29.6	Potentiometric determination.
	98±19	5.0	Ultraviolet and visible spectrophotometric determination.
	97±18	49.0	Ultraviolet spectrophotometric determination.
	97±18	16.4	Other.
glycated hemoglobin (HbA1c) (%, NGSP)** (945 345)†	5.2±0.7	64.9	Immunoassay (latex agglutination turbidimetric immunoassay and so on).
	5.3±0.6	16.9	High Performance Liquid Chromatography (HPLC) determination.
	5.2±0.6	2.0	Enzymatic determination.
	5.2±0.7	16.3	Other.
Urine parameters			
Proteinuria (%) (1 194 283)†	3.8	56.9	Automated dipstick analysis.
	3.9	43.1	Visual dipstick analysis.
Glycosuria (%) (1 195 049)†	2.0	57.2	Automated dipstick analysis.
	2.1	52.8	Visual dipstick analysis.
Fundus oculi examination			
Keith-Wagener hypertensive retinopathy (%) (available n=13 866)	Mild (I) 5.3. Moderate– papilledema (II–IV) 1.0.		Funduscopy.
Scheie hypertensive and sclerotic retinopathy (%) (available n=15 894)	Hypertensive (1–4) 5.4.		Funduscopy.

*Age groups (40–44, 45–49, 50–54, 55–59, 60–64, 65–69 and 70–74 years) are replaced with 42, 47, 52, 57, 62, 67 and 72 years old, respectively, corresponding to the median for each category.
†Available number. No superscript means that the number is completely available (n = 1 209 118).
‡Abdominal obesity: waist circumference ≥85 cm for men or ≥90 cm for women.
§Numbers of risk factors: (1) fasting plasma glucose ≥100 mg/dL and/or HbA1c ≥ 5.6%, (2) triglyceride ≥150 mg/dL and/or high-density lipoprotein cholesterol <40 mg/dL, (3) systolic blood pressure ≥130 mm Hg and/or diastolic blood pressure ≥85 mm Hg and (4) smoking (applicable only for subjects who had at least 1 risk, ranging from 1 to 3).
¶Serum uric acid, creatinine and consequently eGFR are currently unavailable but will become available in the future.
**Almost all subjects (n=1 205 956) had either fasting plasma glucose or HbA1c measured.
BMI, body mass index; eGFR, estimated glomerular filtration rate; HbA1c, glycated hemoglobin; HPLC, high performance liquid chromatography; JSCC, Japan Society of Clinical Chemistry; MetS, metabolic syndrome; NGSP, National Glycohemoglobin Standardization Program.

The standardised 22-item questionnaire created by the MHLW for the health examination check-ups is shown in table 2.

Primary and secondary (minor) outcomes

In a series of studies, we will consider various conditions and aetiologies as primary and secondary (or minor) outcomes (see box 1). However, because unexpected findings are likely to be obtained during these studies and related research topics will be pursued following these findings, we do not restrict the areas of research to be pursued as long as the findings can contribute to or advance specific or general health objectives.

It is noteworthy that subjects are made aware of their health status when their check-up results are complete, and they often receive advice and suggestions from health professionals. Therefore, our cohort studies are not natural history cohort studies by nature. Some proportion of subjects undergo treatments in hospitals, and some receive further health guidance because of the results of their check-ups. Specifically, health guidance for eligible subjects (table 1) aims for the prevention or improvement of mainly cardiometabolic diseases, including MetS. In Japan, medical insurers are required to recommend that individuals at risk of these conditions receive health guidance, although this is not obligatory for the individuals. Health guidance is classified into two categories (intensive and motivational health guidance), depending on the individual's abdominal obesity (waist circumference ≥85 cm for men or ≥90 cm for women) and number of risk factors (table 1). In brief, in cases of intensive health guidance, subjects receive consultation via email, phone or face-to-face sessions for up to 6 months, as has been described in detail elsewhere,[24 33] whereas subjects receiving motivational health guidance do not receive continuous support. Notably, subjects undergoing pharmacotherapy for hypertension, diabetes or dyslipidaemia are excluded, and those aged 65–74 years receive motivational health guidance regardless of their risk profile. In Japan, attendance rates for health guidance have been found to be less than 20%.[24 34]

Statistical analysis

Continuous and categorical variables will be compared between different groups using analysis of variance and χ^2 tests, respectively. Post hoc comparisons between two specific groups will be examined with the Bonferroni, Tukey-Kramer and Dunnett methods, as well as additional χ^2 tests. Paired or trend data will rely on the McNemar, Cochran-Armitage and Mantel-Haenszel test, respectively.

Table 2 Questionnaire on health status and results for 2008

No.	Questionnaire	Answers	Positive response to ① (%)*
	Are you taking following medicines at present?		
1	Medications to reduce blood pressure (1 190 117)†	① Yes ② No	17.1
2	Insulin injection or medications to reduce blood glucose (1 188 900)†	① Yes ② No	3.5
3	Medications to reduce your cholesterol level (1 187 993)†	① Yes ② No	9.4
4	Have you ever been told by the doctor you have a stroke (cerebral haemorrhage, brain infarction and so on) and received treatment for the disease? (978 542)†	① Yes ② No	2.0
5	Have you ever been told by the doctor you have a heart disease (angina pectoris, myocardial infarction and so on) and received treatment for the disease? (978 491)†	① Yes ② No	3.9
6	Have you ever been diagnosed as having a chronic kidney failure and received treatment (dialysis therapy) for the disease? (970 176)†	① Yes ② No	0.3
7	Have you ever been diagnosed as anaemic? (985 060)†	① Yes ② No	10.8
8	Are you a current regular smoker? (1 192 091)† (Here a current regular smoker is to be a person who has smoked a total of 100 or more cigarettes or smoked for 6 months or longer and has been smoking for the last 1 month).	① Yes ② No	25.9 in total. 37.2 in men. 10.8 in women.
9	Have you gained over 10 kg from your weight at 20 years old? (976 268)†	① Yes ② No	35.5
10	Are you in a habit of doing exercise to sweat lightly for over 30 min per session, two times weekly, for over a year? (979 191)†	① Yes ② No	30.9
11	In your daily life do you walk or do any equivalent amount of physical activity more than 1 hour per day? (980 581)†	① Yes ② No	43.6
12	Is your walking speed faster than the speed of corresponds of your age and sex? (964 407)†	① Yes ② No	52.3
13	Have you gained or lost over 3 kg during the last year? (965 421)†	① Yes ② No	22.1
14	How fast do you eat compared with others? (972 294)†	① Faster ② Normal ③ Slower	29.6
15	Do you eat dinner 2 hours before bedtime more than three times per week? (985 764)†	① Yes ② No	28.5
16	Do you eat snacks after supper more than three times per week? (958 850)†	① Yes ② No	12.8
17	Do you skip breakfast more than three times per week? (965 769)†	① Yes ② No	14.6
18	How often do you drink alcohol (sake, distilled spirits, beer, liquor and so)? (989 349)†	① Everyday ② Occasional ③ Hardly drink (cannot drink)	29.2
19	How much do you drink a day, in terms of glasses of refined sake? (A glass [180 mL] of refined sake is equivalent to a medium bottle [500 mL] of beer, 80 mL of shochu [alcohol content 35%], a glass [double, 60 mL] of whiskey and two glasses [240 mL] of wine). (820 231)†	① Less than 180 mL (<23 g ethanol). ② Over 180–<360 mL (23–45 g ethanol). ③ Over 360–< 540 mL (46–68 g ethanol). ④ Over 540 mL (≥69 g ethanol).	56.8
20	Do you feel refreshed after a night's sleep? (973 947)†	① Yes ② No	65.4
21	Do you want to improve your life habits of eating and exercising? (951 484)†	① I don't mean to start. ② I'm going to start in the future (eg, within 6 months). ③ I'm going to start soon (eg, in a month), or I have just started some of them. ⑤ I already started (<6 months ago). ⑥ I already started (≥6 months ago).	28.5
22‡	Are you willing to undergo a health counselling regarding lifestyle modifications if you get the opportunity?	① Yes ② No	–

*Proportions are calculated based on the available numbers for each question.
†Available number of subjects.
‡Question 22 is currently unavailable but will become available in the future.

Box 1 Major and minor outcomes

Major conditions or aetiologies
► Cardiometabolic diseases including type 2 diabetes, hypertension, dyslipidaemia, metabolic syndrome and chronic kidney disease assessed by proteinuria and eGFR.
► Obesity and low body weight (malnutrition) assessed by BMI, and central obesity assessed by WC.
► Hepatic diseases including fatty liver disease assessed by serum AST, ALT and GGT.
► Abnormal eating habits (breakfast skipping, late-night dinner eating, eating fast and night eating).
► Unhealthy lifestyles (smoking, infrequent exercise, heavy alcohol drinking and non-restorative sleep).

Minor conditions or aetiologies
► Hypoglycaemia, hyperfiltrations (high eGFR), hypotension and low uric acid.
► Extremely low and high BMI (eg, $<15.0\,kg/m^2$ and $>40.0\,kg/m^2$).
► Osteoporosis assessed as reduced body height during 6 years.
► Hypertensive and atherosclerotic retinopathies (Keith-Wagener and Scheie classification).
► Physically inactive conditions including reduced walking speed and infrequent exercise.

ALT, alanine aminotransferase; AST, aspartate aminotransferase; BMI, body mass index; eGFR, estimated glomerular filtration rate; GGT, gamma-glutamyl transferase; WC, waist circumference.

Analysis of covariance with general linear model procedures will be used to examine the difference in biochemical variables measured by two or three different methods (eg, low-density lipoprotein cholesterol is measured using three different methods) (table 1), controlling for confounders including age, sex, body weight and various lifestyles. Logistic regression and hazard models will be used to examine the associations or causalities between abnormal levels of measurements and conditions with major and minor outcomes. These methods will yield ORs, risk ratios or HRs, which will be presented along with their 95% CIs. Panel data analysis (including the Hausman test) combining several sets of cross-sectional data will also be conducted. Relevant confounding factors include age, sex, smoking and alcohol consumption, which will be adjusted in the regression analyses. Alternatively, to evaluate or control the differences in backgrounds and various confounders between cases and controls, individuals' propensity scores will be calculated as a variable that unifies all corresponding confounders in the analysis. Propensity score is also considered for a special examination, for instance, hypertensive retinopathy examination, because few subjects underwent such special examination, which yields a bias to be adjusted. The level of health guidance (table 1) and the answer to question 21, which asks about personal intentions to improve eating and exercise habits (table 2), will also be considered as confounding factors, when appropriate.

There are missing data in our study, although this comprises less than 20% of the cases for most parameters and questionnaire items. For categorised age, sex, BMI and WC (table 1), the data are almost complete, even when all of these variables are combined (99.99%). However, combining parameters other than age, sex, BMI and WC can decrease the total available number of subjects, depending on the study's nature and design.

When analysing extremely rare conditions, which might lead to the disclose of the identity of individuals with rare diseases needing treatments in hospital, to prevent the identification of these subjects, we do not describe the number anywhere in the manuscript if this is less than 10, as advised in the MHLW's guidelines.[35]

Statistical analyses will be performed using SAS-Enterprise Guide (SAS-EG V.7.1) in the SAS system, V.9.4. Values of p<0.05 will be considered statistically significant. Because we understand that large data are predisposed to detect the presence of statistical significance, we will use caution in our interpretations and give priority to clinical significance rather than statistical significance in certain clinical areas.

Overall characteristics of subjects

Subjects' characteristics at baseline (2008) are shown in table 1. These findings will vary to some extent in substudies using other cross-sectional data from 2009 to 2014. Men in the first 5 years of their 50s are over-represented in the sample, probably because middle-aged men are more likely than women to work for companies and institutes (ie, insurers), which obligate workers to undergo a check-up. Considering the clinical parameters of BMI, WC, blood pressure, lipids, FPG, HbA1c and MetS, most of the subjects are apparently healthy people with these parameters within normal ranges. The questionnaire results (2008) are shown in table 2, which gives us rough information about the subjects' backgrounds. The smoking rate is higher (25.9%), especially among men (37.2%), compared with other developed countries such as the USA,[36 37] although the smoking rate has been declining in Japan in recent years.[36]

The prevalence of MetS (13.4%), as well as pharmacotherapy for hypertension, diabetes and dyslipidaemia are relatively lower (3.5%–17.1%), compared with other countries.[38 39] However, this does not always mean that the number needed to treat is low, because substantial proportions of subjects likely do not consult a doctor about their poor glycaemic control. A Japanese national survey conducted from 2005 to 2009 found that, among people with diabetes, a substantial proportion (about 38%) have left their poor glycaemic conditions as they are, without seeking treatment.[40] In our study, the extent of this issue remains unknown without a detailed investigation of FPG and HbA1c.

Concerning eating habits, which play an important role in metabolism and nutrition, the percentage of subjects who habitually skip breakfast is lower (14.6%) than the percentage who eat dinner 2 hours before bedtime (28.5%), which is consistent with a previous study.[41] This suggests that the latter group may be more troublesome in terms of unfavourable lifestyle habits that are linked

to cardiovascular diseases, because a close association between eating dinner late at night and skipping breakfast has been reported in a community-based epidemiological study[41]; eating dinner late at night can lead to skipping breakfast the next morning. Acknowledging the relationship with sleep, we term these behaviours 'unfavorable eating habits around sleep' (UEHAS). In previous studies,[41 42] eating dinner late at night together with skipping breakfast—a combination representative of UEHAS—was significantly associated with MetS, proteinuria and atrial fibrillation.

Body weight substantially influences the incidence and development of cardiometabolic diseases as well as general health.[43 44] However, features and aetiologies at both extremes of BMI, a fundamental index of weight considering height, body adiposity, nutritional status and health, are poorly understood. For rare conditions in malnutrition, for instance, the percentage of subjects with an extremely low body weight (BMI <15.0 kg/m^2, a criterion for high mortality[45]) is very small in the cross-sectional dataset in 2008 (0.1%), but the observational number in this dataset is large (n=1217, data not shown), which is not ignorable and may be enough to conduct proper statistical analyses. Likewise, extremely high body weight (BMI ≥40.0 kg/m^2, a criterion for class III obesity[46]), is also very small in percentage (0.07%), but there are 805 observations for this group in this dataset.

Throughout the study presented here, we expect to obtain a wide range of novel observations, enabling us to confirm indeterminate previous findings, especially in terms of cardiometabolic disease. In addition, this study will likely reveal underlying aetiologies that have been overlooked because they are rare or minor cases in small clinical studies.

Among our research team members, one person (Nakajima) has previously been involved in a similar large study, consisting of approximately 100 000 people living in Saitama Prefecture (population=7.3 million people), which is also located near Tokyo[47] (figure 1). This previous study and its substudies were launched in 2011, and to date, multiple findings have been reported from these studies. Although a study of 100 000 people is generally considered 'large', this sample size sometimes proved to be inadequate for stratification analysis because of small observational numbers for particular groups when several stratification variables were combined—for instance, age group, BMI category and diabetes status.[48] This is the main reason we chose to begin a new study using the check-up data for an extraordinarily large sample—over one million for a cross-sectional study, which is more than 10 times higher than this previous study.[47]

BMI roughly reflects nutritional status, including excess energy accumulation or malnutrition. In the last decade, increased percentages of people have been found at both extremes of BMI (ie, in the nutritional states of malnutrition or obesity), worldwide, especially among children. This coexistence of undernutrition and obesity or nutrition-related non-communicable disease has been termed

the 'double burden of malnutrition'.[49–51] In our study, subjects' characteristics described above suggest that the double burden of malnutrition can exist even among the middle-aged Japanese population, although the proportions are smaller compared with other developing countries.[49]

Our study has several strengths. First, in terms of community-based epidemiological research, the sample is so large that precise results concerning the means of parameters and 'normal' values, although these are standard parameters, can be obtained,[52] even when subjects are classified into categories such as sex, age group, smoking status and certain morbidities (eg, obesity or diabetes). Therefore, it may be possible to conduct similar analyses to produce novel results in other population studies with very large databases that allow for multiple classifications. A second strength of this study is that it may be possible to use big data to evaluate minor or rare conditions and aetiologies that are commonly overlooked, neglected or unfeasible to analyse in clinical studies, particularly those with small samples.[14] This analysis may contribute to case studies instead of only to the field of public health. Finally, identical measurements and assessments of anthropometric indices, blood pressure and urinalysis are performed across multiple years in people living in similar environment and the same healthcare system.

Some limitations to this study should also be mentioned. First, the variations in parameters are restricted, and parameters for specific diseases are not included, because the check-ups are conducted for general health and the prevention of common diseases, especially lifestyle-related diseases such type 2 diabetes and MetS. Second, people younger than 40 years and those aged over 74 years are not enrolled in this study. Lifestyle choices made when people are younger may contribute to the incidence of morbidities in middle age, and lifestyles and clinical biochemistry levels in middle age can influence the incidence and severity of cardiovascular diseases and health damage in the later life. Unfortunately, comparison with younger and older people is unfeasible, so a seamless analysis over the life course is impossible in this study. Third, although cohort analysis using this dataset is possible, at 6 years, the period is relatively short, which may hamper the ability to uncover the latent relationships and underlying mechanisms between the parameters used and the predicted outcomes. Durations of 10 years or even several decades may be needed to clarify the latent causality between suspected factors and outcomes.[53] Finally, to date, there is no comprehensive and concise definition of big data.[54] It is therefore unclear whether the term 'big data' applies to our database. Big data is commonly characterised by volume, variety, velocity and veracity,[18 19 54] and some of these terms (volume and veracity) may be applicable to our database. However, a larger database including the latest datasets and longer durations of observation may be required to have the characteristics of 'big data', which enable researchers to use emerging

analysis tools, including artificial intelligence techniques such as machine learning.

In our composite study, we expect to obtain a wide range of novel findings and to confirm indeterminate previous findings, with multidisciplinary applications, especially in terms of cardiometabolic disease. We also expect this work to provide new perspectives for human health promotion and disease prevention.

Author affiliations
[1]School of Nutrition and Dietetics, Faculty of Health and Social Services, Kanagawa University of Human Services, Yokosuka, Japan
[2]Department of Endocrinology and Diabetes, Saitama Medical Center, Saitama Medical University, Kawagoe, Japan
[3]Department of nutrition, St. Marianna University School of Medicine, Kawasaki, Japan
[4]Graduate School of Health Care Sciences, Jikei Institute, Osaka-shi, Japan

REFERENCES

1. Ma RC, Chan JC. Type 2 diabetes in East Asians: similarities and differences with populations in Europe and the United States. *Ann N Y Acad Sci* 2013;1281:64–91.
2. Lu Y, Hajifathalian K, Ezzati M, *et al*. Metabolic mediators of the effects of body-mass index, overweight, and obesity on coronary heart disease and stroke: a pooled analysis of 97 prospective cohorts with 1·8 million participants. *Lancet* 2014;383:970–83.
3. Arai H, Ouchi Y, Toba K, *et al*. Japan as the front-runner of super-aged societies: Perspectives from medicine and medical care in Japan. *Geriatr Gerontol Int* 2015;15:673–87.
4. Suzuki T. Health status of older adults living in the community in Japan: recent changes and significance in the super-aged society. *Geriatr Gerontol Int* 2018;18:667–77.
5. Hayashi F, Takimoto H, Yoshita K, *et al*. Perceived body size and desire for thinness of young Japanese women: a population-based survey. *Br J Nutr* 2006;96:1154–62.
6. Sugawara A, Saito K, Sato M, *et al*. Thinness in Japanese young women. *Epidemiology* 2009;20:464–5.
7. Ministry of Health, Labour and Welfare. National Nutrition Survey. 2013 http://www.mhlw.go.jp/file/04-Houdouhappyou-10904750-Kenkoukyoku-Gantaisakukenkouzoushinka/0000068070.pdf (cited 10th Sep 2015).
8. Fried LP, Tangen CM, Walston J, *et al*. Cardiovascular Health Study Collaborative Research Group. Frailty in older adults: evidence for a phenotype. *J Gerontol A Biol Sci Med Sci* 2001;56:M146–56.
9. Kido Y. The Issue of nutrition in an aging society. *J Nutr Sci Vitaminol* 2015;61:S176–7.
10. Arai H. [Sarcopenia, frailty and osteoporosis]. *Clin Calcium* 2017;27:1279–85.
11. Wouterse B, Huisman M, Meijboom BR, *et al*. The effect of trends in health and longevity on health services use by older adults. *BMC Health Serv Res* 2015;15:574.
12. de Magalhães JP, Stevens M, Thornton D. The business of anti-aging science. *Trends Biotechnol* 2017;35:1062–73.
13. Ministry of Health, "Labour, and Welfare. Health examination and guidance program for Japanese Adults. 2008 http://www.mhlw.go.jp/bunya/shakaihosho/iryouseido01/info02a.html
14. Kaplan RM, Chambers DA, Glasgow RE. Big data and large sample size: a cautionary note on the potential for bias. *Clin Transl Sci* 2014;7:342–6.
15. Jagodnik KM, Koplev S, Jenkins SL, *et al*. Developing a framework for digital objects in the big data to knowledge (bd2k) commons: report from the commons framework pilots workshop. *J Biomed Inform* 2017;71:49–57.
16. Nizielski SE, Lechner PS, Croniger CM, *et al*. Animal models for studying the genetic basis of metabolic regulation. *J Nutr* 1996;126:2697–708.
17. Petters RM, Sommer JR. Transgenic animals as models for human disease. *Transgenic Res* 2000;9:347–51.
18. Bellazzi R. Big data and biomedical informatics: a challenging opportunity. *Yearb Med Inform* 2014;9:08–13.
19. Lee CH, Yoon HJ. Medical big data: promise and challenges. *Kidney Res Clin Pract* 2017;36:3–11.
20. Conceptual diagram. Provision of medical-related data to a third party. http://www.mhlw.go.jp/english/topics/2011eq/workers/tepco/lhc/ref10_20140604.pdf
21. Kubo S, Noda T, Myojin T, *et al*. The need and key points for patient matching in clinical studies using the National Database of Health Insurance Claims and Specific Health Checkups of Japan (NDB). *Japanese Journal of Health & Research* 2017;38:11–19.
22. http://www.kuhs.ac.jp/kenkyuu/2017121400013/ Japanese.
23. Health Research Involving Human Subjects. Ethical guidelines for medical and health research involving human Subjects. http://www.mhlw.go.jp/file/06-Seisakujouhou-10600000 Daijinkanboukouseikagakuka/0000080278.pdf
24. Health Care Expenditures. The final interim report by the work group for studying the effects of the specific health checkups and specific health guidance on health care expenditures. http://www.mhlw.go.jp/file/05-Shingikai-12401000-Hokenkyoku-Soumuka/0000123428.pdf
25. https://www.mhlw.go.jp/file/06-Seisakujouhou-10900000-Kenkoukyoku/00_3.pdf
26. The definition and diagnostic criteria of metabolic syndrome. *Nippon Naika Gakkai Zasshi*;200.5:794–809.
27. Grundy SM, Brewer HB, Cleeman JI, *et al*. Definition of metabolic syndrome: Report of the National Heart, Lung, and Blood Institute/American Heart Association conference on scientific issues related to definition. *Circulation* 2004;109:433–8.
28. Rodríguez-Ortiz D, Reyes-Pérez A, León P, *et al*. Assessment of two different diagnostic guidelines criteria (National Cholesterol Education Adult Treatment Panel III [ATP III] and International Diabetes Federation [IDF]) for the evaluation of metabolic syndrome remission in a longitudinal cohort of patients undergoing Roux-en-Y gastric bypass. *Surgery* 2016;159:1121–8.
29. Kashiwagi A, Kasuga M, Araki E, *et al*. International clinical harmonization of glycated hemoglobin in Japan: from japan diabetes society to national glycohemoglobin standardization program values. *J Diabetes Investig* 2012;3:39–40.
30. Matsuo S, Imai E, Horio M, *et al*. Revised equations for estimated GFR from serum creatinine in Japan. *Am J Kidney Dis* 2009;53:982–92.
31. Keith NM, Wagener HP, Barker NW. Some different types of essential hypertension: their course and prognosis. *Am J Med Sci* 1974;268:336–45.
32. Sairenchi T, Iso H, Yamagishi K, *et al*. Mild retinopathy is a risk factor for cardiovascular mortality in Japanese with and without hypertension: the Ibaraki Prefectural Health Study. *Circulation* 2011;124:2502–11.
33. Tsushita K, S Hosler A, Miura K, *et al*. Rationale and descriptive analysis of specific health guidance: the nationwide lifestyle intervention program targeting metabolic syndrome in Japan. *J Atheroscler Thromb* 2018;25:308–22.
34. Ichikawa D, Saito T, Oyama H. Impact of predicting health-guidance candidates using massive health check-up data: a data-driven analysis. *Int J Med Inform* 2017;106:32–6.
35. https://www.mhlw.go.jp/file/05-Shingikai-12401000-Hokenkyoku-Soumuka/0000135460.pdf

36. Health risks. Health risks - Daily smokers - OECD Data. https://data.oecd.org/healthrisk/daily-smokers.htm

37. Jamal A, Agaku IT, O'Connor E, *et al*. Current cigarette smoking among adults--United States, 2005-2013. *MMWR Morb Mortal Wkly Rep* 2014;63:1108–12.

38. Falkner B, Cossrow ND. Prevalence of metabolic syndrome and obesity-associated hypertension in the racial ethnic minorities of the United States. *Curr Hypertens Rep* 2014;16:449.

39. Engin A. The definition and prevalence of obesity and metabolic syndrome. *Adv Exp Med Biol* 2017;960:1–17.

40. Goto M, Goto A, Ikeda N, *et al*. Factors associated with untreated diabetes: analysis of data from 20,496 participants in the Japanese National Health and Nutrition Survey. *PLoS One* 2015;10:e0118749.

41. Kutsuma A, Nakajima K, Suwa K. Potential association between breakfast skipping and concomitant late-night-dinner eating with metabolic syndrome and proteinuria in the Japanese Population. *Scientifica* 2014;2014:253581:1–9.

42. Nakajima K, Suwa K, Oda E. Atrial fibrillation may be prevalent in individuals who report late-night dinner eating and concomitant breakfast skipping, a complex abnormal eating behavior around sleep. *Int J Cardiol* 2014;177:1124–6.

43. Ul-Haq Z, Mackay DF, Fenwick E, *et al*. Meta-analysis of the association between body mass index and health-related quality of life among adults, assessed by the SF-36. *Obesity* 2013;21:E322–E327.

44. Corbin LJ, Timpson NJ. Body mass index: Has epidemiology started to break down causal contributions to health and disease? *Obesity* 2016;24:1630–8.

45. Kotler DP. Cachexia. *Ann Intern Med* 2000;133:622–34.

46. Poirier P, Cornier MA, Mazzone T, *et al*. Bariatric surgery and cardiovascular risk factors: a scientific statement from the American Heart Association. *Circulation* 2011;123:1683–701.

47. Muneyuki T, Suwa K, Oshida H, *et al*. Design of the saitama cardiometabolic disease and organ impairment study (scdois): a multidisciplinary observational epidemiological study. *Open J Endocr Metab Dis* 2013;03:144–56.

48. Nakajima K, Suwa K. Excess body weight affects HbA1c progression irrespective of baseline HbA1c levels in Japanese individuals: a longitudinal retrospective study. *Endocr Res* 2015;40:63–9.

49. Abdullah A. The double burden of undernutrition and overnutrition in developing countries: an update. *Curr Obes Rep* 2015;4:337–49.

50. World Health Organization. *The double burden of malnutrition: policy brief*. Geneva: World Health Organization, 2017.

51. Demaio AR, Branca F. Decade of action on nutrition: our window to act on the double burden of malnutrition. *BMJ Glob Health* 2018;3:e000492.

52. Manrai AK, Patel CJ, Ioannidis JPA. In the era of precision medicine and big data, who is normal? *JAMA* 2018;319:1981–2.

53. Bell JA, Hamer M, Sabia S, *et al*. The natural course of healthy obesity over 20 years. *J Am Coll Cardiol* 2015;65:101–2.

54. Mayer-Schönberger V, Ingelsson E. Big Data and medicine: a big deal? *J Intern Med* 2018;283:418–29.