Chapter 2
Design, Implementation, and Management of Biobank Studies

Zhengming Chen

Contents

Abstract Prospective biobank studies are required for reliable assessment of the importance of both genetic and non-genetic causes of disease and their complex interplay in disease aetiology. Unlike case–control studies involving cases of particular diseases, prospective studies typically include healthy individuals recruited from the general population, with their health status monitored for several years or decades in order to identify a sufficient number of incident cases to reliably assess their associations with particular risk exposures. However, because only a small proportion of the study participants will develop any particular disease each year, such studies typically need to include several hundreds of thousands of participants and to continue follow-up for an extended duration in order to accrue sufficient numbers of diseased cases for reliable analyses. Hence, meticulous planning and preparation are needed to ensure that the appropriate scientific, organisational, and ethical frameworks are in place before establishing such biobank studies. This chapter addresses the basic principles and practical issues that should be considered in planning, designing, and conducting large-scale population-based prospective

Z. Chen (✉)
Big Data Institute Building, Nuffield Department of Population Health, Old Road Campus,
University of Oxford, Oxford, UK
e-mail: zhengming.chen@ndph.ox.ac.uk

Z. Chen (ed.), *Population Biobank Studies: A Practical Guide*,
https://doi.org/10.1007/978-981-15-7666-9_2

biobank studies including the collection and storage of biological samples. The general principles and approaches required for such studies are also applicable to other studies in different settings or using different designs (e.g., cross-sectional surveys and case–control studies).

Keywords Cohort studies · Biobanks · Protocol · IT · Record linkage · Data management · Quality assurance · Ethics · Governance

Abbreviations

API	Application Programming Interface
CKB	China Kadoorie Biobank
COPD	Chronic obstructive pulmonary disease
CT	Computed tomography
DNA	Deoxyribonucleic acid
ECG	Electrocardiogram
EDTA	Ethylenediaminetetraacetic acid
GDPR	General data protection regulation
ID	Identifier
ISAB	International Scientific Advisory Board
IT	Information technology
MRI	Magnetic resonance imaging
NCD	Non-communicable chronic disease
PI	Principal investigator
RNA	Ribonucleic acid
SBP	Systolic blood pressure
SOP	Standard operation procedures
URS	User requirement specification

2.1 Introduction

Large prospective biobank studies are essential for assessing the role of lifestyle, environmental and genetic factors, and their complex interplay, in disease aetiology. In contrast with case–control studies, information on exposures in prospective studies is collected before the onset of disease, which minimises the risk of reverse causality bias (Hennekens and Buring 1987). The information on associations with incident cases of diseases in individuals who are exposed to certain risk factors are compared with those who are not exposed in the same population, which minimises the risk of selection biases, in addition to enabling the appropriate temporal sequence between exposures and disease outcomes. Prospective studies can simultaneously examine the associations of many different disease outcomes with particular exposures (e.g., tobacco smoking), or, in studies with stored biological samples with multiple biochemical, genetic, or novel multi-omics biomarkers. However, prospective studies are expensive and time-consuming to conduct. Moreover, there are major challenges in conducting prospective biobank studies, including strategies to achieve complete long-term follow-up, and obtain reliable classification of disease outcomes. Additional challenges include the need to collect updated measures of exposure status at periodic intervals in all or random subsets of study participants. Therefore, in planning the study careful attention should be paid not just to the initial recruitment of participants but also to long-term follow-up of their health outcomes. Planning should also include consideration of study design and data collection methods, quality assurance of exposure and disease outcome data, study organisation, and management, in addition to ethics and governance issues. The key for ensuring success lies not in planning a perfect study, but rather in planning the most appropriate, reliable, and sustainable study given the practical constraints of resources, time, and capacity. Thus, a successful biobank study requires an appropriate balance between the science of the theoretical and the art of the practical.

2.2 Plan of Investigation

The first step in planning a successful prospective study is to ask why such a study is needed by formulating the research aims and objectives in a research proposal. After finalising a research proposal, a detailed study plan should be developed. The research proposal and study plan should be guided by a careful literature review of the existing evidence and extensive consultation with the scientific community, followed by pilot studies to ensure that the planned procedures and systems are fit for purpose. In developing a detailed study plan, careful consideration should be given to both scientific and practical issues related to selection of the study population, sample size and sampling methods, assessment of risk exposures and disease outcomes, in addition to ethical and cost implications (Grimes and Schulz 2002).

2.2.1 Study Population

Depending on the research objectives, risk exposures, and health outcomes of interest, the population under investigation may vary by age, gender, occupation, and certain other factors. For example, pre-menopausal women should be targeted in studies to investigate the long-term health effects of use of the contraceptive pill, while newborn babies should be recruited into birth cohorts to assess the role of prenatal factors for child health or development. However, the recruitment of middle-aged men and women is more appropriate for prospective studies of the aetiology of chronic non-communicable diseases (NCDs). In defining specific selection criteria for the study population, due consideration should be given to their perceived exposure levels and health status, anticipated future disease rates, ease of recruitment (including any communication problems), and likely mobility over time.

For example, it would not be appropriate now to target doctors in the UK to study the health effects of smoking because hardly any doctors currently smoke, in contrast to the situation 40–50 years ago. Likewise, young adults (e.g., aged <35 years) may not be appropriate for inclusion in prospective studies because they are difficult to recruit, to trace long-term, and have very low immediate risk of developing NCDs. On the other hand, old people (e.g., >75 years) may not be considered suitable, as they may have many existing health problems that could greatly distort risk exposure assessments (e.g., low body mass index [BMI] due to existing diseases, low plasma LDL-cholesterol level due to current use of statins), resulting in misleading associations of exposures with diseases. For these reasons, most prospective studies tend to select middle-aged adults (e.g., aged 35–70 years) who tend to be relatively healthy, geographically stable, and have a reasonably high risk of developing NCDs in the near future.

2.2.2 Sample Size

Every study should have the appropriate statistical power with the ability to generate reliable answers to the proposed research questions. In prospective studies, it is certainly true that the bigger the sample size, the better the study, provided that it is feasible and that the quality of the data collection and completeness of long-term follow-up can also be maintained. The desired sample size can be readily estimated using online statistical programmes, which take account of both statistical factors (e.g., planned study power, perceived effect size of exposures, and level of statistical significance) and other factors (e.g., prevalence of exposures, loss to follow-up, disease event rates). However, the sample size estimation can only be indicative in prospective studies involving many different exposures and many different diseases. Moreover, such algorithms do not consider one of the most important factors in determining the sample size and that is the level of financial support that can be secured. Theoretically, the study design influences the costs, but in practice the converse is often true. In planning the desired sample size, it is frequently necessary to do the exercise in reverse, and consider how the sample size can be maximised given the likely available resources. Invariably, there is a need for a trade-off between sample size and complexity of data to be collected. Typically, there is a risk of making a study too complicated (e.g., by including an excessive number of questions or measurements), often at the cost of a reduced sample size. It should also be recognised that even a really large prospective study involving 0.5 million participants (e.g., UK Biobank (Sudlow et al. 2015) or China Kadoorie Biobank [CKB] (Chen et al. 2011)) may still not be big enough for studies of rare diseases (or other less common conditions) or to quantify reliably the effects of exposure on common diseases in specific population subgroups (e.g., by age, or levels of other exposures). Prolonged follow-up of individual studies will be needed in order to accrue a sufficiently large number of disease cases, complemented by efforts to combine data in meta-analyses of findings from multiple similar studies.

2.2.3 Sampling Methods

In certain epidemiological studies (e.g., cross-sectional surveys of the prevalence of risk exposures), it may be desirable that the study population selected is representative of the general or target population, in which case random sampling is usually required to ensure that all members of the population have an equal chance of being selected and there is no systematic non-response. In prospective studies, however, such approaches may not be feasible or necessary, and use of random sampling approaches will greatly increase the costs and complexity in organisation and long-term follow-up for disease outcomes. Prospective studies of non-representative cohorts of individuals with heterogeneity in risk exposures can still generate reliable evidence about the associations of particular risk factors with disease outcomes that are widely generalizable (Chen et al. 2020a, b). For example, findings from the British Doctors Study, initiated in 1951 and including periodic resurveys of smoking habits in each decade for over five decades, demonstrated that smoking is a major cause of lung cancer and >20 other diseases and the findings remain relevant not only for doctors but also for the worldwide population (Doll et al. 2004).

In prospective studies, certain communities or subgroups of the general population (e.g., physicians, nurses, and civil servants) are typically selected as the target population to maximise the response rate and minimise the loss to follow-up. Depending on the planned sample size and anticipated response rate, all or a subset of the target population within a catchment area or community is typically invited to enrol in a study. In large prospective studies involving multiple regions or localities, the selection of study sites should also consider geographic location, patterns of major disease rates and risk exposures, levels of economic development, and estimated population mobility, in addition to the local infrastructure (including quality of existing death or disease report systems, availability of courier service for sample shipment) and long-term commitment to the project. Given that participation in the study is typically on a voluntary basis and a proportion of individuals invited will not respond for various reasons, it is necessary to estimate the response rate and likely reasons for non-participation in a random subset of non-responders, so that their impact on study planning, future data analyses, and generalizability of the findings can be reliably assessed.

2.2.4 Exposure Measures

The types and ranges of risk exposures (and potential confounding factors) to be assessed in prospective studies may vary depending on the study objectives, perceived importance and relevance of different exposures for different diseases, and available resources. For research into the aetiology of NCDs, they would generally cover several distinctive aspects, including: (1) demographic and socio-economic factors (e.g., age, sex, marital status, education, and income); (2) lifestyle factors (e.g., diet, tobacco smoking, alcohol drinking, and physical activity); (3) reproductive patterns (e.g., age of menarche, parity, and breast feeding); (4) occupational or environmental factors (e.g., exposure to indoor and ambient air pollution); (5) physical and biochemical characteristics (e.g., height, adiposity, lung function, hand grip strength, blood pressure, liver and renal function, blood lipid and glucose levels); (6) personal and family medical history and use of certain medications (e.g., antihypertensive treatment, lipid lowering treatment, and hormonal replacement therapy among post-menopausal women); (7) sleep, cognitive and psychological state; and (8) genetic factors. To ensure data quality and completeness, information on risk exposures should generally be collected using appropriately designed questionnaires and carefully planned physical measurements in addition to the collection, storage, and laboratory assays of biological samples.

2.2.4.1 Questionnaires

The study questionnaires are the key documents used to collect relevant information on exposures from all participants. It is likely that individual biobank studies have different research interests, and hence study-specific questionnaires are often required. Questionnaires may be self-administered or administered by interviewers. Self-administered questionnaire are cost-effective and less susceptible to interviewer bias and may be more appropriate to collect information on sensitive issues (e.g., sexual behaviour or finances), and can be conducted by mail or using the Internet. On the other hand, the non-response rates may be high and answers to certain questions may be incomplete. Moreover, in a population with a high illiteracy rate, it may not be feasible to use self-administered questionnaires. Depending on the study population, survey procedures and questions included, both approaches can sometime be combined in the same study (e.g., UK Biobank) for collecting data on different types of questions.

To ensure consistency and facilitate future data analysis, each question should have, where possible, a closed-response format (with the exception of numerical answers such as number of cigarettes smoked), in which the respondent is provided with a list of pre-determined response options. Open-ended questions involving non-numerical text messages can elicit a more detailed response, but the responses may vary greatly, which will require more effort to extract and encode relevant information for data analysis. Such open-ended questions may be useful in the initial pilot study to help select and refine a list of pre-determined response options. For closed-response questions, the list of pre-determined responses should include all possible options. For certain questions, the participant may not know the answer for various reasons (e.g., birth weight or exposures during childhood), which should be permitted (e.g., by entering a symbol "#" or having a category "Don't Know") to differentiate them from missing values (i.e., unanswered). Each question should be simple, factual, and properly worded to avoid any ambiguity. Moreover, it should cover one dimension, with comprehensive and mutually exclusive choices of

Table 2.1 Comparison of computer-based versus paper-based questionnaire interviews

	Computer-based	Paper-based
Technical support	Complex	Simple
Initial cost	High	Low
Delivery speed	Slow	Fast
Training of staff	Easy	Difficult
Ease of use	Easy	Difficult
Flexibility (e.g., sub-form)	High	Low
Quality control	Easy	Difficult
Data quality	Good	Poor
Data release	Fast	Slow

answers. Furthermore, it should be phrased in a way that will not influence the likely response in one direction or another.

Typically, there is a tendency to include too many questions, which may greatly increase the cost and reduce compliance by participants. Information collected in a questionnaire should be based on and limited to the objectives of the study. For prospective studies with very broad objectives, it may be necessary to develop certain criteria to prioritise selection of questions to be considered during the planning phase. These may include: (1) the perceived strength of evidence about hypotheses of exposure–disease relationships; (2) the anticipated prevalence (e.g., at least 15%) in the population; (3) the public health importance of the relevant condition in particular populations; (4) the likely importance of factors that might act as confounders or sources of bias; (5) the reliability and validity of questionnaire measures; and (6) the availability of alternate sources of information about the factor (e.g., medical records, physical measurements). Where possible, it is preferable to adapt multiple questions from previously validated questionnaires used in other studies. The questionnaires should always be tested in pilot studies prior to inclusion in the main survey to assess their feasibility, comprehension and acceptability of each question, time taken to complete each of them, and response rates.

Where possible, computerised direct data entry methods should be used in preference to conventional paper questionnaires. These will not only facilitate training and improve the efficiency of data collection processes (e.g., avoiding printing, transport and storage of questionnaires, and manual data punching) but also allow internal quality (e.g., avoidance of any missing values) and consistency checks, automated coding, immediate access for ongoing central monitoring and audit, and rapid data release for research purposes (Table 2.1).

2.2.4.2 Physical Measurements

With advent of rapid technology development, a wide range of physical measurements can be considered in prospective studies. They can be used to improve our understating of disease aetiology (e.g., blood pressure, body mass index, bio-impedance), risk prediction (e.g., hand grip strength, lung function), and early diagnosis (e.g., ECG, bone density, liver fibro-scan, carotid intima-medial thickness and plaque, and CT/MRI scans) for many different conditions. They can also allow a more objective and continuous assessment of certain risk exposures (e.g., accelerometer for physical activity and sleeping patterns). Again, given there are many possible options, the selection of physical measurements should be based on the study objectives, with careful consideration of their perceived scientific value, relevance for particular conditions, reliability of the data collected, and available resources.

In planning the range of measurements and selecting from different device models for particular measurements, it is also necessary to consider certain practical issues, such as time taken for the measurement, size and likely mobility of the device, ease of use, environment required for the test (e.g., private room vs open space), and any discomfort that may be caused to participants. As for the cost involved, apart from the initial purchasing cost, it is also important to consider associated costs including operator requirements (e.g., technician vs clinical specialist), service contracts, and consumables required. For each measurement considered, it is important to have a quality assurance framework to ensure data quality and integrity. This should involve maintenance, calibration, training, monitoring, and data transfer to IT systems. Specifically, the operation of each device should be managed (or controlled) by study computers on site through an API (Application Programming Interface). The API can be provided by device manufacturers and/or developed purposely by the study team with technical support from manufacturers, and will enable direct entry of certain personal details (e.g., study IDs) that can be linked to the measurement data and instant transfer of data from the device to study computers (see Chap. 7).

2.2.4.3 Collection of Biological Samples

In prospective studies, biological samples collected can generate the most important information about determinants, prevention, early detection and treatment of many diseases. Depending on the study objectives, a wide range of biological samples can be considered both from participants (e.g., blood, urine, saliva/buccal cells, faeces, hair and nails, placental tissue, cord blood, breast milk) and the living environment (e.g., air, water, soil). In general the aim of the sample collection and procedures involved should be “future proof”, i.e., to allow the widest possible range of assays that could plausibly be envisaged in the future given the current knowledge and available resources. For the reasons described in Table 2.2, blood and urine samples should always be prioritised in any prospective studies. Other types of samples might allow measurements of certain factors not covered by blood or urine (e.g., hair and nails for assessing exposure to environmental heavy metals, and faeces for gut microbiome), but they may be difficult to collect and process (e.g., faeces), and may not accurately reflect exposure at personal levels (e.g., ambient air pollution), and will add significant additional costs for collection, processing, shipment, and long-term storage.

Table 2.2 Rationale for collecting blood and urine samples in biobank studies

Sample type	Reasons for consideration
Blood	• A variety of fractions: plasma, serum, white cell, red cells, peripheral blood lymphocytes • Wide range of biomolecules: DNA, RNA, proteins, small molecules • Wide coverage of physiological functions: genome, proteome, and metabolome, haematological parameters • Suitable for a wide range of assay technologies • Ease and low cost of collection
Urine	• Wide range of biomolecules: proteins, analytes (including pharmaceuticals) • Wide coverage of physiological functions: proteome and metabolome (including gut microbiome) • Suitable for many assay technologies • Ease and low cost of collection

For collection of blood and urine samples, there are a wide variety of collection tubes with different preservatives and additives. Careful review of preservatives and anticoagulants in such tubes is important when planning the collection and future assays, as certain anticoagulants are recommended for some but contraindicated for other assays (Elliott and Peakman 2008). For example, blood samples collected into EDTA-containing tubes have optimal DNA yields and hence are ideal for DNA-based assays, but may be unsuitable for assays of potassium, calcium, magnesium, and zinc because of chelation of such ions. Likewise, heparin-stabilised blood affects T-cell proliferation assays and heparin binds to many proteins. In most cases, the selection of additives is a compromise and if a choice has to be made, then EDTA-containing tubes for sample collection are considered optimal because they can allow valid measurements of genetic markers (using DNA-containing buffy coat) and a very wide range of biomarkers (using red cells and plasma), using both conventional and novel multi-omics assay platforms. Depending on the available resources, the types of assay to be conducted immediately or planned in the future, and the long-term storage facilities, the sample volume to be collected from each participant should be planned carefully. Importantly, many modern omics assay platforms only require a small sample volume, involving about 100 μl of plasma, for assays of many hundreds or even thousands of non-genetic biomarkers simultaneously. Similarly, buffy coat in a 10 mL EDTA blood sample should yield enough purified DNA for undertaking a range of genetic assays, including whole genome sequencing.

The sample collection tubes should be properly labelled with barcodes that can be linked to participant's original study ID. Samples collected at the assessment centres or survey clinics should be kept chilled, usually refrigerated at 4 °C, and then transported and processed at a local or central laboratory with as little time delay as possible (ideally within 12 h). The blood sample can be separated into different fractions after centrifugation (e.g., plasma, red cells, white cell "buffy" coat), which are usually aliquoted, either manually or using an automated working station, into multiple smaller storage tubes suitable for long-term cryopreservation. Throughout the process, the reliability of sample tracking and identification is essential, which often requires support of robust IT and quality assurance systems (see Chaps. 4 and 7).

2.2.5 Long-Term Follow-Up

The value of a prospective study depends not only on its ability to obtain detailed baseline data and samples from a large number of individuals but also on detailed follow-up of their health status, including death, disease occurrences, and changes in lifestyles, and other risk exposures over time.

2.2.5.1 Periodic Resurveys

The risk exposures measured at the initial baseline survey are subject to measurement error, biological variation, and long-term changes over time, which can lead to "regression dilution bias" (Clarke et al. 1999) when assessing associations of such exposures with disease outcomes occurring many years or decades after recording such measurements. This regression dilution bias causes substantial underestimation of the strength of long-term "usual" levels of such risk factors with disease outcomes, but can be corrected for by estimating the extent of the within-person variation, usually by conducting periodic resurveys of random samples of surviving participants every few years. In small studies with a few thousand participants, it may be possible to resurvey all surviving participants. In large studies involving hundreds of thousands of participants, it may only be feasible to conduct periodic resurveys of random samples of 5–10% of surviving participants.

In addition to repeating identical data collection as at the baseline, certain enhancements can also be considered to address future research questions, including new questions, new samples, new measurements that become feasible, or improved measures of certain risk exposures (e.g., accelerometer measures of physical activity and sleep patterns). If a high proportion of the study population have access to the Internet, then certain questionnaire-based resurveys (e.g., dietary or cognitive assessments) can be repeated more regularly and involve all or a large proportion of the participants. To minimise selection bias, every effort should be made to achieve a high response rate. After the first resurvey, the subsequent resurveys can involve a high proportion of the same participants who were selected initially, which will help to provide a more reliable assessment of time trends and changes with increasing age of the main risk exposures.

2.2.5.2 Disease Outcomes

Cause-specific mortality is the most widely used health outcome in prospective studies and should be prioritised. Where feasible, it is also important to consider other health outcomes (e.g., disease incidence, episodes of hospitalisation), which will greatly increase the range of diseases that can be studied (e.g., non-fatal diseases) and improve the study power and accuracy of disease diagnosis. This may also facilitate research into areas that are not conventionally feasible in prospective studies, such as the natural history and management of specific diseases (Chen et al. 2020a, b) The information about health outcomes can be obtained through re-contact of participants (i.e., *active* follow-up) which has been widely used in many prospective studies previously. Although this approach can obtain certain health information that may not be well represented in record linkage data including repeat exposure measures, the response rate may be low (typically <70%) and costs may be prohibitively high, especially in large studies with regular contact. Moreover, the information obtained directly from participants about disease diagnosis is usually less complete and reliable. The most efficient and reliable way of obtaining health outcome data is through *passive* follow-up, i.e., linkage with available datasets including death and cancer registries, health insurance claim databases, or primary health care records. In certain populations, it may also be possible to obtain linkage with histopathological records using hospital tissue repositories. Such linkages can be achieved electronically using certain matching algorithms or unique personal identification numbers collected at the baseline survey, which will enable the cost-effective follow-up of the whole cohort in a timely manner (see Chap. 5).

To facilitate follow-up and minimise potential loss to follow-up of participants over time, the study areas should be carefully selected at the planning stage to ensure that the population in the catchment areas is relatively stable and that the available health record systems are adequate. In areas without established death and disease registries, alternative strategies for follow-up should be carefully planned and piloted before launching the main study. Moreover, all the individuals considered should be permanent residents within the catchment area and have their personal details (e.g., national ID number, telephone number, and email address) carefully and confidentially recorded during the survey. Once recruited, the follow-up for health outcomes of study participants should start immediately without waiting until after completion of the whole baseline survey, which may be many years later. To ensure the completeness of follow-up and reliability of the disease diagnosis, it is necessary to cross-check and validate outcome data collected from different sources. Moreover, for certain major health outcomes (e.g., stroke, cancer, COPD), independent investigations are also needed to verify and sub-phenotype the reported disease diagnoses through retrieval and review of medical records (see Chap. 6).

2.3 Ethical and Legal Considerations

In most countries, formal ethical approval for biobank studies will be needed from relevant institutions or other organisations. Increasingly, the ethical committees will not only review consent procedures and related documents but also consider the validity of the proposed study design (e.g., sample size and selection bias) in addition to issues related to data protection and confidentiality. In general there are four areas of interest: (1) legal requirements regarding data collection and storage, especially when they may carry certain risks or they are related to genetic and medical information; (2) confidentiality of data provided to the study by the participant; (3) access to data held on the study population by other sources and in particular, their medical records; (4) sharing of the study data with other researchers. There are specific legal and ethical requirements for investigators to protect and maintain confidentiality of the data collected, which are fairly complex, and this is an increasingly important issue in epidemiological studies. A general framework should be considered carefully according to the official guidelines issued by relevant bodies in the country concerned (e.g., the UK official Data Protection Act 1998, UK Human Tissue Act 2004, and the EU General Data Protection Regulation [GDPR] 2018). These guidelines provide legal frameworks and a set of "principles", which must be adhered to by the study investigators.

Consent is necessary for all research involving human subjects, which protects both the participants and the study. It is mandatory in most countries to obtain written consent from participants in prospective studies for a number of reasons. These include: (1) survey procedures may involve certain risks; (2) investigations undertaken (e.g., ultrasound or CT scan) may uncover previously unrecognised conditions that may require further intervention; (3) the need to obtain information (e.g., medical records) from a third party; (4) the need for long-term storage of biological samples for unspecified research use in future; and (5) protection of personal information collected. It is evident that most people participate in the study for the purpose of supporting academic research in an altruistic manner. In general the research institutes and principal investigators have legal responsibilities for the proper custody and use of both biological samples and data collected from participants. As for the use of biological samples, the level of consent (i.e., narrow or broad) may vary depending on the study goals and local rules and regulations. Where possible, consent should be kept broad and future proof in order to maximise the potential of samples collected.

To facilitate the formal consent process, the study leaflet or invitation letter should provide clear, accurate, and complete information about the study. In general they should cover the following points: (1) a clear statement that the study is for research purposes, and that participation is voluntary and non-participation will not disadvantage them in any ways; (2) the exact nature of the study, including the study purpose, organisation, official approvals obtained, procedures involved, and any potential risks they may incur; (3) indicating why they were chosen (e.g., at random), whether they will be given any test results and how long the study will last; (4) a statement indicating even if the subject agrees to participate they may opt out for certain items (e.g., without providing biological samples) and can withdraw at any time without giving any specific reasons, although individuals who have expressed such wishes before joining the study should not be encouraged to participate; and (5) a clear statement indicating how personal information and data provided will be protected and used. The formal consent form to be signed by participants should contain a clear statement that participants have been given full information and have the opportunity to raise and discuss any issues with staff. The consent form should also list separate data and sample collection items, for which specific consent by participants may be required. With respect to incidental findings of previously unrecognised conditions, the action to be taken will depend on the nature and severity of the problem, its natural history, and the availability of any effective intervention. In certain cases, the information may need to be referred back to participant's doctor for further consultation and examination.

2.4 Study Protocol

Once developed, a study plan should be recorded as a written protocol to provide overall guidelines for the conduct and day-to-day running of the study. The protocol should describe the rationale for the study, its main objectives and the methodology used, and should describe each essential component of the study, from eligibility of the participants, sample size and sampling schemes, through types and methods of data collection and follow-up, to study organisation, ethics, budgets and governance. The protocol is also an essential component of a research proposal for funding applications and for obtaining necessary ethical approvals from relevant institutions and regulatory agencies.

The study protocol should be developed after a careful and thorough review of existing literature and appropriate consultation with colleagues, collaborators, and experts in the fields. If necessary, pilot studies should be undertaken to test and refine the study design, detailed work plan, and data collection tools (e.g., questionnaires). Once a study protocol has been developed and approved, and the study has started and progressed, it should be adhered to strictly, with any subsequent changes kept minimal and carefully documented with the file reference number and release date. In general the study protocol for prospective studies should cover the following aspects: (1) Title; (2) Project summary; (3) Rationale and background; (4) Study purposes and objectives; (5) Study design and plan, including study population, sample size, recruitment, data and sample collection, and follow-up; (6) Data management and statistical analyses; (7) Study organisation; (8) Ethics and governance; and (9) Budget and timelines. To help develop the study protocol, operational procedures and quality assurance framework, various working groups should be established with shared objectives and coordinated efforts and approaches.

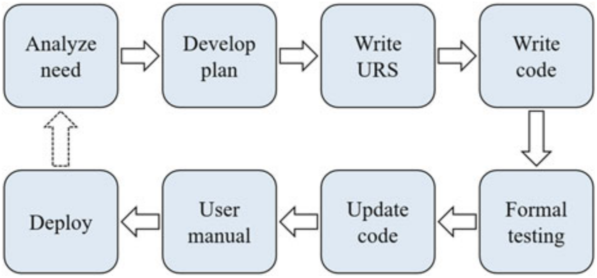


Fig. 2.1 Standard process for developing IT software in biobank studies

2.5 IT Infrastructure and Systems

IT support is one of the most important cornerstones of a successful biobank study. Given the highly specialised nature of biobank studies, it is unlikely that many off-the-shelf software packages will be readily available to support particular studies. Hence, a range of bespoke IT systems will need to be developed to manage all aspects of the study activities. The IT system should not just cover data collection (e.g., questionnaire interview, physical measurements, and sample collection), but also cover the management of staff, data, assets and consumables as well as quality control and study monitoring. If developed and implemented successfully, they will help ensure and maintain consistency, traceability, timeliness, and quality of the data collected over time and across different centres and staff, while at the same time reducing costs and unnecessary workload for project staff. For example, many physical measurement devices need to be calibrated and serviced on a regular basis. Instead of relying on study staff to remember, various schedules can be incorporated into a study IT system that can automatically monitor the usage or performance of specific devices and send out requests according to pre-determined roles (e.g., number of tests done, consistency of the performance over time). Similarly, the IT systems for data collection can also incorporate specific functions to facilitate monitoring and quality control. For example, the laptop-based questionnaire can have an audio recording function to record all or part of the interview, which can be reviewed and checked centrally.

Depending on the resources, local capacity, technical needs, and timeline, the study IT systems can be developed in-house (i.e., directly employ IT development staff) or outsourced (i.e., pay another organisation, usually commercial, to develop). Each approach has its strengths and limitations. Although the initial cost may be higher and time delay longer, the long-term benefits of in-house development in terms of ease and cost of maintenance, upgrading, quality control and system integration would greatly outweigh the initial shortcomings, which is the development model adopted in the CKB. Irrespective of how they are developed, the IT industry standard for developing procedures and methodologies should be followed (Fig. 2.1), including preparation of detailed User Requirement Specification (URS) documents, and formal testing. Throughout the lifecycle of development, study

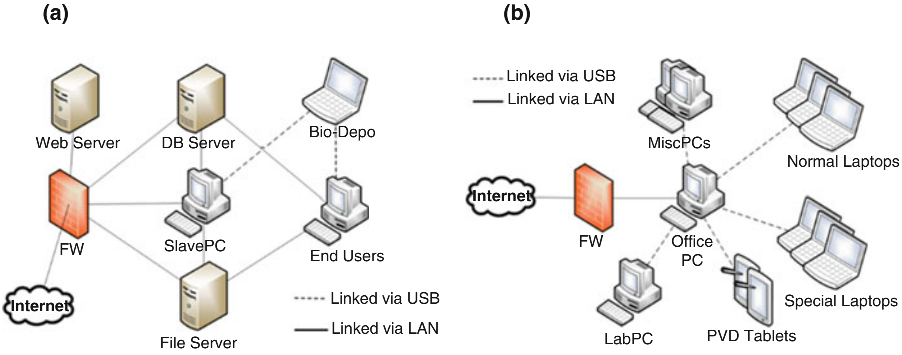


Fig. 2.2 IT network and infrastructure at national and regional study centres in CKB. (**a**) National Coordinating Centre, (**b**) Regional Study Centre

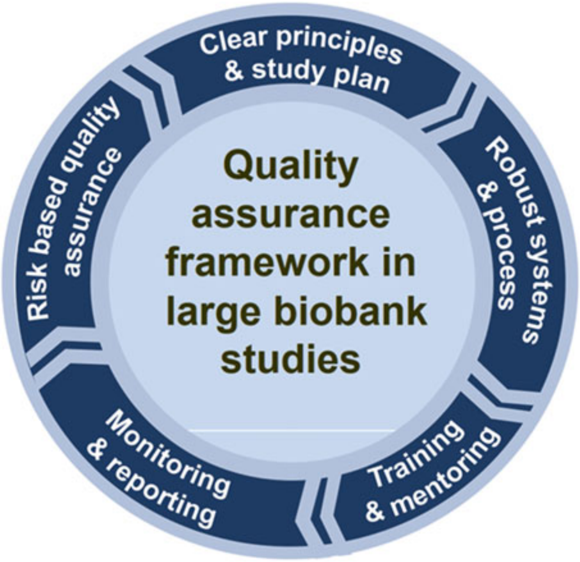
investigators should be closely involved in defining and specifying requirements and functionalities, including preparation of a URS for each system (see Chap. 7).

IT hardware devices that may be considered in biobank studies will vary, ranging from mobile phones, desktop computers, laptops, tablets, to servers, or even large cloud-based supercomputing and storage facilities. Very often different types of IT devices will be needed in the same study to meet different requirements and settings. Apart from hardware and software, other factors should also be carefully considered when planning and developing study IT infrastructure, including Internet connection, firewall, data size, regulations, and local IT support staff. Depending on the study need and settings, there may be very different arrangement for IT infrastructure at national and local study centres (see Fig. 2.2 and Chap. 7).

2.6 Quality Assurance Framework

Quality assurance refers to the planning, policies, training, procedures, and actions necessary to ensure that the quality, integrity, and ethical standards of the study are being maintained and enhanced during the course of the study. Given the complexity and length of prospective biobank studies, a quality assurance framework should be developed to provide an evidence-based, robust, coordinated, and cost-effective approach to quality assurance. It should be implemented across various stages of study, from planning and designing, through development of operational procedures, training, and field work, to monitoring and improvement (see Fig. 2.3). Apart from study design, training, and development of robust systems and process that are covered in different sections, careful attention should be paid to study documentation, monitoring, and data management. Where possible, IT systems should be incorporated to facilitate the process.

Fig. 2.3 Diagram of quality assurance framework in large biobank studies


Clear principles & study plan
Robust systems & process
Training & mentoring
Monitoring & reporting
Risk based quality assurance
Quality assurance framework in large biobank studies

2.6.1 Pilot Study and Documentation

Before launching the main study, it is essential to undertake several pilot studies, not only to test questionnaires, methods for recording physical measurements, and IT systems, but also to assess recruitment strategies, staff needs and training requirements, practical procedures, and logistics in addition to scheduling and coordination. Moreover, to assure a uniform, consistent, and standardised approach to carrying out the study with good quality control, Standard Operation Procedures (SOPs) should be developed to provide detailed and specific instruction to the investigators. The SOPs should cover not only data collection (e.g., interview, physical measurements, and sample collection) but also data management, study logistics (e.g., supply, sample shipment), and organisation (e.g., staff training, assessment centre). All the study equipment and devices should be properly documented in a central inventory, with regular calibration and servicing according to the manufacturer's recommended schedule.

2.6.2 Management of Data and Information

Apart from data collection, specific procedures, IT systems, and data management plans should also be developed to manage data transfer, processing, integration, access, and use to ensure that the security, confidentiality, traceability, consistency, and integrity of the data can be properly maintained through the life course of the study. All databases should be stored and handled securely, with different levels of authorised access across all study locations and with proper separation of personal details from any study data collected for research use. A central data repository should be established with regular and comprehensive backups and change logs (see Chap. 8). The decommissioning of any study IT devices should also be handled carefully and securely. Prior to equipment disposal, all confidential information should be securely erased and physically destroyed. For purposes of future auditing, a mirror copy of all the data held (e.g., in survey laptops, office desktop computers, or servers) should be made and stored in a central data repository. A record of the destruction of devices and data should be logged for future reference.

2.6.3 Study Monitoring

Study monitoring should be undertaken on a regular basis by the coordinating centre, using a combination of computer review of data and periodic on-site monitoring visits. The computerised data review should focus on recruitment rates, missing data or biological samples, data quality (e.g., the number of outliers, difference between two measures), visit and sample processing time, and performance of centre (e.g., wastage of consumables, service and calibration of devices). After accumulation of a reasonably large number of participants in each centre and by different staff, it is prudent to undertake statistical monitoring of the data collected, for example, by examining the distribution of the exposure data (e.g., height, blood pressure), prevalence of certain risk exposures (e.g., tobacco smoking) over time and by different centres and by different staff in order to detect any outliers, inconsistencies, or potentially fraudulent data. Similar monitoring should also be extended to long-term follow-up (see Chap. 5). Any issues or problems identified during this continuous review of the data should be followed up by telephone conference or a site visit by staff from the study coordinating centre. Figure 2.4 illustrates findings of routine statistical monitoring in CKB for standing height and systolic blood pressure (SBP), showing consistency in measured values over time for the former, but not for the latter. Further investigation revealed that the seasonal variations in SBP, evident in all ten study areas, were not a data quality issue, but driven primarily by changes in ambient temperature. This led to several publications, with important public health and clinical implications (Lewington et al. 2012; Yang et al. 2015).

2.7 Study Assessment Centre

In prospective studies, study assessment centres are typically needed in order to enrol a large number of participants from local communities. Depending on the requirements, the assessment centres can be located either in established clinical facilities, serviced commercial office space, or local public premises (e.g., school, village hall). Whichever type of assessment centre is selected, it should be located

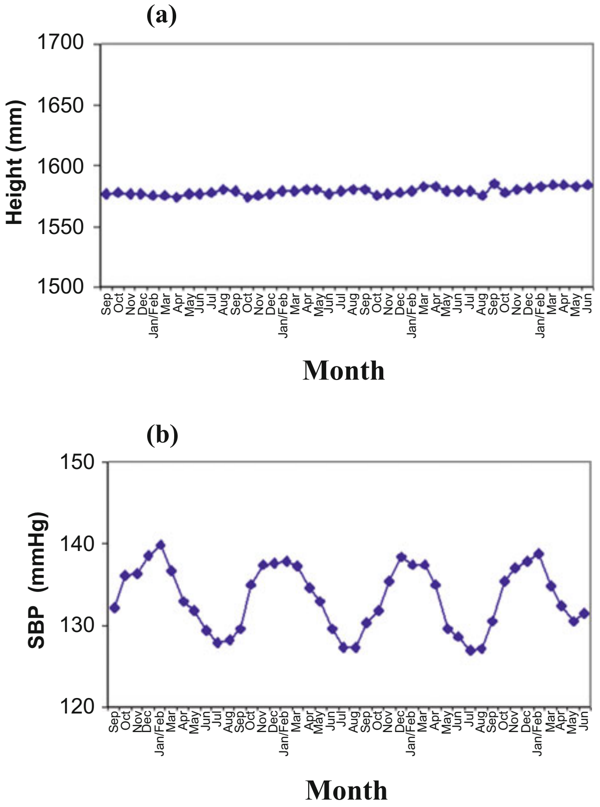


Fig. 2.4 Statistical monitoring of measured standing height and blood pressure in CKB. (**a**) Standing height, (**b**) Systolic blood pressure

conveniently within the study area, have good transport links, and have a default level of services (e.g., lavatories, electricity, and water). The area covered by individual assessment centres should enable recruitment of a sufficient number of potentially eligible study participants within a specific time period (e.g., 2–4 months), taking into account the likely response rate and estimated daily recruitment rate.

2.7.1 Centre Configuration

Standard requirement specifications including the likely floor plan and survey flow should be developed to ensure that each assessment centre can be configured to meet the study needs (e.g., enough power sockets, Internet connection, secure space for a small server, and quiet rooms). The arrangement of different assessment stations should be carefully planned to ensure there are little or no bottlenecks in flow and certain measurements are conducted in the appropriate sequence where possible (e.g., measuring blood pressure before lung function, which requires strenuous effort leading to increased blood pressure). For questionnaire interviews or certain physical measurements that may require more time, it is necessary to have multiple stations in order to reduce likely bottlenecks. In areas where a reliable power supply cannot be guaranteed, it is also necessary to have a mobile power generator as a backup. An assessment centre equipment specification should be constructed based on the study plan. A set of equipment should be procured for each of the centres running in parallel, which should be inventoried and regularly serviced and calibrated. In case of breakdown, backup equipment should be held centrally or on site. The use of consumables in each operating assessment centre should be monitored carefully, with a small buffer stock held locally to compensate for greater-than-projected demand and with the main supply managed by the study coordinating centre.

2.7.2 Survey Team and Training

The field survey team should be properly structured and staffed according to the study plan, planned recruitment rate, number of stations, anticipated workload and time taken for each station. The team should have a senior manager or coordinator who will have overall responsibilities and report back directly to study coordinating centre. For each post, a detailed job description and requirements including qualifications, and relevant skills and experiences should be considered. All the staff recruited should receive appropriate training, involving not only by formal lectures but also practical training, testing, and simulated data collection ("dry runs"). In case of holiday or sick leave, certain staff should also be trained to have dual roles as backup. After completion of formal training, the field work should start without any undue delay, operating perhaps at half capacity initially to enable staff to become familiar with the procedures. The initial phase of the field work should be supervised and supported by senior members of the steering committee and study coordinating centres, with daily review meetings to discuss outstanding issues and areas for further improvement. Staff involved in field work recruitment should be mentored throughout the study period.

2.7.3 Recruitment of Participants

Where feasible, the eligible participants should be identified in advance through national or local population-based registers (e.g., National Health Service or public security records). The formal invitation letters can be generated centrally or locally and then be delivered by post or manually by local staff or community leaders. To increase awareness and participation rates, publicity campaigns and social mobilisation might be necessary, involving mass media as well as community meetings. When invited, potential participants should be given provisional appointments, with clear instructions on the essential documents that they should bring with

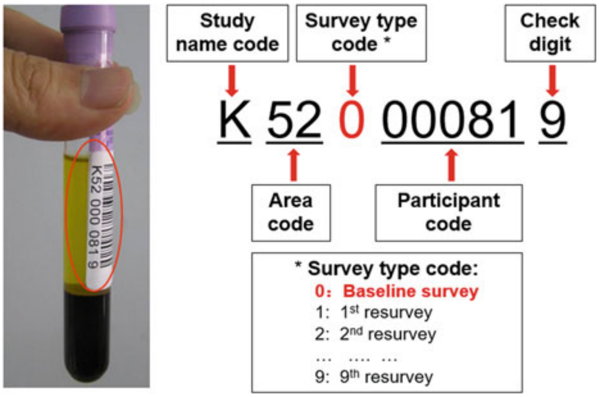


Fig. 2.5 Example of study ID used in CKB

them (e.g., appointment card, national ID card) to the assessment centre. When an individual arrives at the assessment centre, they will need to provide formal consent first and then move through a series of assessment stations. Each consenting participant should be allocated a unique study ID number (see Fig. 2.5) that is linked securely to their personal details and study and sample data. The study ID number, usually in a form of barcode number, can be printed on the consent form, or stored in a USB memory key allocated to each participant at the assessment centre. At each assessment station, the study ID number should be carefully checked and recorded, usually through a barcode reader, to ensure reliable linkage and integration of the data collected. Towards the end of assessment, the participant may be given a formal report with all measurement results, which can then be discussed with a medically qualified physician in the assessment centre (see Chap. 3).

2.8 Central Biobank Infrastructure

To ensure their long-term security, biological samples collected should be stored centrally, perhaps at separate locations. The storage temperatures may vary depending on the material itself, its “robustness” and anticipated length of storage (ISBER 2008; Elliott and Peakman 2008). Ideally, most samples (e.g., plasma, serum, buffy coat, urine, peripheral blood circulating cells) should be stored below the re-crystallisation temperature of pure water at −130 °C, for which liquid nitrogen tanks (vapour face or liquid face) will be preferred. A working archive (i.e., short-term storage) of basically the same set of samples can be stored at a higher temperature in −80 °C freezers. Genomic DNA, especially when amplified, should be stored at −20 °C. The central sample storage facilities should be managed by trained staff, with appropriate alarm systems, backup electricity and power generators, which should be tested on a regular basis. Moreover, all the samples checked in or retrieved should be carefully documented and tracked using sample management systems (see Chap. 4). For large biobank studies involving millions of aliquots, it is

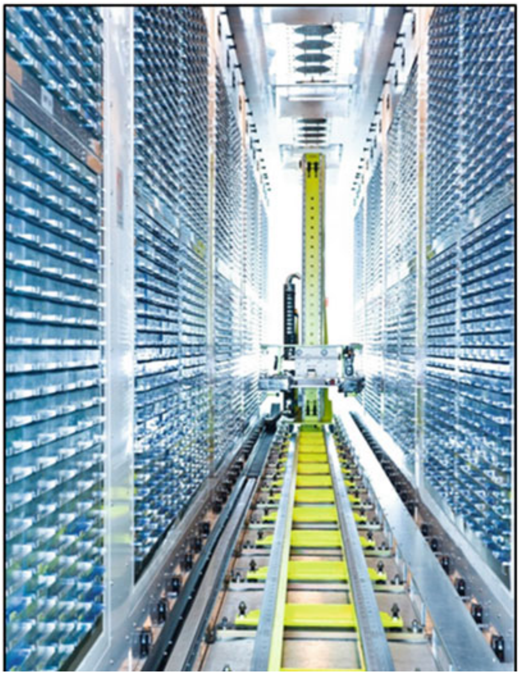


Fig. 2.6 Automated sample storage and management system in UK Biobank (re-use with permission from Peakman and Elliott 2010)

necessary to install a fully automated sample storage and management system, as has been used successfully in the UK Biobank (Fig. 2.6).

2.9 Study Organisation and Oversight

After developing a detailed research protocol, operational plan, and quality assurance framework, the study should be implemented with scientific rigour to ensure that the study protocol is being adhered to, and that the research is conducted in accordance with established procedures and ethical standards. In addition, meticulous and detailed records of all data and information should be maintained and properly documented, and methods of data collection used in a consistent way by different staff and over time. To achieve these, effective study organisation, oversight, and management are essential.

2.9.1 Steering Committee

The Steering Committee is responsible for the overall leadership and management of the study. It will provide scientific input into the development of the study protocol, and on the direction and scientific objectives of the project. It will also oversee the operation of the project, including recruitment of study participants, the sample collection, processing and archiving strategy, the development of approaches for long-term follow-up of participants' health outcomes, reviewing and approving of study budgets, and plans for funding raising. Moreover, the Steering Committee will review and approve study governance and other policy documents, external collaborative projects and membership for the International Scientific Advisory Board (ISAB).

2.9.2 Coordinating Centres

Depending on the study plan, organisational structures, and numbers of survey sites and locations, separate central and local coordinating centres may be needed to coordinate and organise the study. For both centres, there should be proper provision of adequate space and staff, with clearly defined roles and operational structures. In general the central coordinating centre will be responsible for study planning, obtaining ethical and regulatory approvals, development of SOPs and computer software, organising training and collaborators' meeting, purchase of study equipment and devices, preparing and distributing study materials to survey sites (e.g., information leaflets, sample collection kits), management and storage of data and biological samples, monitoring and auditing study progress, administration of budget and contracts, responding to technical, medical, and administrative queries, and preparation of progress report to funders and the steering committee.

The local coordinating office in each survey site will be chiefly responsible for the reliable conduct of the field survey. This should involve obtaining local approval, the identification of study sites and participants, establishment of the survey team and assessment centres, organisation of field surveys, processing and shipment of biological samples, and dealing with any inquiries that the study participants may raise. If the long-term follow-up for disease outcomes needs to be carried out locally, then the local coordinating office should also be responsible for obtaining formal approval and negotiate contractual and cost issues with local government agencies for accessing health records, and for undertaking long-term follow-up of health outcomes in addition to verification and adjudication of disease diagnoses.

2.9.3 Scientific Advisory Board

For large prospective studies with very broad objectives, it is necessary to establish an ISAB, to provide advice to the study PI and steering committee on the scientific direction, long-term strategy and operations. It may also review progress and achievements against the agreed objectives and also review future plans and provide advice on fund-raising activities and prioritisation of research projects to be undertaken.

For each committee or board, it would be helpful to develop a formal charter, defining detailed roles, responsibilities, scope of activities, length of service, time schedules, and appointment procedures. In addition, other high-level governing body or council oversight may be required, if the study is set up initially as a national resource involving multiple institutes and funders (e.g., UK Biobank).

2.10 Summary

This chapter provides a high-level overview of scientific and practical considerations for establishment of large prospective biobank studies. Many of the issues discussed and possible solutions suggested reflect to a large extent the thorough processes involved in setting up the large CKB study of >0.5 million participants who were recruited during 2004–2008 from 10 geographically diverse urban and rural areas across China. The future chapters will provide more detailed descriptions of several specific areas of work related to biobank studies, including field work, sample collection and handling, long-term follow-up and disease event adjudication, development of IT systems, and data management. It is intended that the main focus will be on general principles and practical approaches so that they can be applied to many other future studies in different settings or using different designs.

References

Chen Z, Chen J, Collins R, Guo Y, Peto R, Wu F, Li L. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. Int J Epidemiol. 2011;40:1652–66.

Chen Z, Emberson J, Collins R. Strategic need for large prospective studies in different populations. JAMA. 2020a;323:309–10.

Chen Y, Wright N, Guo Y, Turnbull I, Kartsonaki C, Yang L, Bian Z, Pei P, Pan D, Zhang Y, Qin H, Wang Y, Lv J, Liu M, Hao Z, Wang Y, Yu C, Peto R, Collins R, Li L, Clarke R, Chen ZM. Mortality and recurrent vascular events after first incident stroke: a 9-year community-based study of 0.5 million Chinese adults. Lancet Glob Health. 2020b;8:e580–e90.

Clarke R, Shipley M, Lewington S, Youngman L, Collins R, Marmot M, Peto R. Underestimation of risk associations due to regression dilution in long-term follow-up of prospective studies. Am J Epidemiol. 1999;150:341–53.

Doll R, Peto R, Boreham J, Sutherland I. Mortality in relation to smoking: 50 years' observations on male British doctors. Br Med J. 2004;328:1519–33.

Elliott P, Peakman TC. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. Int J Epidemiol. 2008;37:234–44.

Grimes DA, Schulz KF. Cohort studies: marching towards outcomes. Lancet. 2002;359:341–5.

Hennekens CH, Buring JE. Epidemiology in medicine. Boston: Little, Brown and Co.; 1987.

International Society for Biological and Environmental Repositories (ISBER). 2008 best practices for repositories: collection, storages, retrieval and distribution of biological materials for research. Cell Preserv Technol. 2008;6:3–58.

Lewington S, Li LM, Sherliker P, Millwood I, Guo Y, Collins R, Chen JS, Whitlock G, Lacey B, Yang L, Peto R, Chen ZM. Seasonal variation in blood pressure and its relationship with outdoor temperature in 500,000 adults in 10 areas of China, the China Kadoorie Biobank. J Hypertens. 2012;30:1383–91.

Peakman T, Elliott P. Current standards for the storage of human samples in biobanks. Genome Med. 2010;2:72.

Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins R. 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(3):e1001779. https://doi.org/10.1371/journal.pmed.1001779.

Yang L, Li LM, Lewington S, Guo Y, Sherliker P, Bian Z, Collins R, Peto R, Liu Y, Yang R, Zhang YR, Li GC, Liu SM, Chen ZM. Outdoor temperature, blood pressure and cardiovascular disease mortality among 23,000 individuals with diagnosed cardiovascular diseases from China. Eur Heart J. 2015;36:1178–85.