

Zhengming Chen *Editor*

# Population Biobank Studies: A Practical Guide

 Springer

# Population Biobank Studies: A Practical Guide

Zhengming Chen  
Editor

# Population Biobank Studies: A Practical Guide

 Springer

*Editor*

Zhengming Chen  
Nuffield Department of Population Health  
University of Oxford  
Oxford, Oxfordshire, UK

ISBN 978-981-15-7665-2      ISBN 978-981-15-7666-9 (eBook)  
<https://doi.org/10.1007/978-981-15-7666-9>

© Springer Nature Singapore Pte Ltd. 2020

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.  
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

# Foreword

## *A large study is not just a small study made larger*

There can be few books into which the scientific editor has put more decades of epidemiological preparation than Zhengming Chen has into this book or at least into the study that it describes. Now Professor of Epidemiology at Oxford, he began doing epidemiological studies in China more than 30 years ago. He has continued ever since, working on studies in China from a UK base and building in the process the largest collaboration in the world between Chinese and Western randomised and observational studies of population health. Each new study has been considerably larger than its predecessors.

Eventually, he and the team he had built up at Oxford were conducting, jointly with co-principal investigators and colleagues in China, rigorously randomised trials with several tens of thousands of participants of widely practicable treatments for common diseases. They were also conducting larger and larger observational epidemiological studies, culminating 15 years ago, as described in this book, in what was the world's largest blood-based biobank study, with samples stored from half a million apparently healthy adults all over China, with electronic linkage to all deaths and to virtually all hospital treatment via the newly introduced nationwide health insurance scheme.

Fortuitously, this was just the moment when information technology, sample storage and retrieval, health record linkage, assay technology (genetic and non-genetic), and statistical methods had improved so much that, with detailed attention both to the organisation and to the science, a really large biobank study could succeed.

Equally fortuitously, a substantial one-off grant from the Kadoorie Charitable Foundation in Hong Kong, structural support from the Disease Surveillance Points system, and long-term support from Oxford University Departmental infrastructure gave Professor Chen and his colleagues the initial freedom to concentrate on optimising the planning, conduct, and maintenance of this, the first major biobank study of the new century.

Because it had been planned and executed so carefully and successfully, the China Kadoorie Biobank Study (CKB, which recruited 500,000 Chinese adults in 2004–2008) provided an influential model when, in the mid-2000s, a complete redesign was undertaken of the UK Biobank Study (UKB, which then successfully recruited 500,000 UK adults in 2008–2010).

CKB is now maintained by long-term support from major funding agencies in Beijing and London (with continued support from the Oxford's Nuffield Department of Population Health, which hosts both CKB and UKB), but the initial freedom offered by the early financial supporters of both studies was crucial to the careful planning and piloting that underlay their eventual efficiency and success.

Although these two biobank studies grew out of the twentieth-century tradition of prospective studies, in their methods and size they went far beyond it. In turn, they have provided methodological examples of successful use of twenty-first-century techniques that have inspired and influenced biobank studies elsewhere.

Currently, in a welcome development, all the major biobank studies in the world are communicating with each other, sharing methods, ideas, data, and results. This book can become part of the process of sharing methods, both with other studies and with the future. This matters, for as Rory Collins, chief executive officer and “onlie begetter” of UK Biobank has observed, a large study is not just a small study made larger.

In recent years vast numbers of scientific (and unscientific) articles have been written about the promise and problems of big biobank studies. Still, however, too little has been written by the few who have actually made such studies work reliably and productively. The interconnected problems, which need to be planned against, are partly organisational, partly technical (assay methods are improving so rapidly that it is often better to procrastinate, waiting for big decreases in price and increases in sensitivity), and partly statistical.

Statistical traps are laid by regression dilution (which can be avoided by appropriate use of periodic resurveys of a subsample of the study population), by unduly fine subgroup analyses, by random variation in measurements (as adjustment for imperfectly measured confounding factors can leave highly significant residual confounding), and by misleading relationships between different imperfectly measured factors that cannot be adequately resolved by multiple regression or by the recently fashionable “directed acyclic graphs” (DAGs). Another major problem can be reverse causality, but this can often be adequately dealt with by exclusion of those who already had disease at study entry (and, for some associations, exclusion of the first few years of follow-up).

Finally, as the randomised and the observational methods used so successfully in studies of physical disease get used increasingly widely in studies of mental disease, education, criminology, social policy, international development and many other issues, understanding the real problems that have been encountered and overcome in large prospective studies of the physical disease may be of increasingly wide interest.

Although a few highly entertaining moments have had to remain censored (and the Editor has East Asian flushing syndrome, so they cannot be elicited by alcohol), what remains is an account of a remarkable and influential study, relevant to the conduct and interpretation of all major prospective studies over the next decade or two.

University of Oxford, Oxford, UK

Sir Richard Peto FRS

# Preface

Common chronic diseases of adults have their roots in lifestyles, social factors, chronic infections, environment, in addition to genes. Last several decades have witnessed significant progress in our understanding of the main avoidable causes of common diseases, driven in part by development in epidemiological research, especially prospective cohort studies. The introduction of prospective studies can be traced to the late 1940s or early 1950s when several landmark studies were established, including the British Doctor Studies, the Framingham Study, and the Study of the Japanese Atomic Bomb Survivors. Such studies were undertaken to address major pressing public health concerns at the time, such as the health consequences of smoking, the causes of the escalating burden of heart disease, and the risks from radiation. Some of these studies continued into the twenty-first century and have greatly improved disease prevention, risk prediction, and treatment.

While the key principles of prospective studies have remained unaltered since their introduction, recent advances in information technology, exposure assessment, molecular biology, and genetics have greatly transformed the ways by which prospective studies are conducted in the twenty-first century. Contemporary blood-based prospective studies, now often referred to as “population biobank studies”, tend to be extremely large and complex. They typically involve extensive collection of exposures and disease outcomes and long-term storage of biological samples for future large-scale (or cohort-wide), hypothesis-free, multi-omics assays. The need for large sample sizes, which is essential for assessing modest but biologically important associations, requires feasible strategies to facilitate recruitment, minimise loss to follow-up, accurately classify disease outcomes, optimise the use of limited biological samples, and maintain secure long-term storage of biological material and data. Given the challenges, the key to success lies, perhaps not in planning for a perfect study, but rather in planning the most reliable, sustainable, and future proof study within the practical constraints of available resources and capacity.

In my day-to-day work as an Epidemiologist at the University of Oxford over the past three decades, I have been involved in teaching graduate students and become increasingly aware of the paucity of textbooks that address the contemporary



practice of epidemiology in the era of “Big Data”. Many of the available epidemiology textbooks focus on theoretical approaches, with varying degrees of statistical complexity. Moreover, many textbooks including “practical guide” or “handbook” in their title tend to focus on the data processing or analytic aspects of epidemiological studies, rather than addressing practical issues or the limitations of some of the traditional approaches. For example, studies that rely on paper-based questionnaires incur a disproportionate amount of time to check, clean, and process data before allowing researchers to undertake statistical analyses of such data. Another important issue that has not been fully appreciated is the importance and challenges of conduct of very large blood-based studies (e.g., 0.5 million or more participants in a single study) or the methodology required to establish and manage such studies reliably and cost-effectively across many geographically diverse areas.

This book aims to bridge the gap between traditional and contemporary epidemiology and describes the key components of prospective biobank studies, including the procedures and quality assurance frameworks for optimal design, conduct, and management of such studies. The book involves eight chapters including: (1) introduction (overview of principles and methods); (2) design and management of biobank studies; (3) organisation and management of field work; (4) management of biological samples; (5) long-term follow-up for health outcomes; (6) verification and adjudication of disease outcomes; (7) development and application of IT systems; and (8) data management and curation. Individual chapters can be read separately or together. Some issues have been presented from distinct, albeit interrelated, perspectives in several chapters, including questionnaire design, standard operating procedures, ethics and regulatory approval, biological samples, best practice in software development, and data protection and sharing. Using examples mainly from the China Kadoorie Biobank, the book provides many practical case studies of state-of-the-art, cost-effective, and scalable methods necessary for establishing large biobanks in different settings.

The authors are members of a multidisciplinary team of epidemiologists, clinicians, geneticists, software engineers, and laboratory and data scientists currently working on the China Kadoorie Biobank at the University of Oxford. With the first-hand experience in the planning, design, conduct, and management of contemporary biobank studies, the authors have endeavoured to share their experience with their readers. Even as a practical guide book, the chief emphasis is not to offer simple solutions but to provide an in-depth analysis of various practical issues that are likely to be encountered. The book could serve as a useful reference book to those who wish to study epidemiology or undertake population health research. Individuals who would like to understand more about the discipline may also find it informative. It should also be of general relevance for researchers in other fields. I hope that this book will also stimulate further development of large population- and hospital-based biobank studies across different populations.

# Commentary Remarks on “Population Biobank Studies: A Practical Guide” Edited by Chen

The main focus of this book by Chen and his colleagues is on prospective studies of associations between risk factors assessed at baseline in the participants (including stored biological samples) and the health outcomes that occur during their subsequent follow-up.

Typically, when planning such studies, the focus is on the scientific purposes and on appropriate ways to address them. By contrast, the authors consider not only key aspects of study design required to address particular scientific aims, but also—based on their considerable practical experience—how to deliver the desired study effectively.

Several large prospective studies already exist, but there is still a need to establish additional studies in carefully selected populations in different parts of the world. This book will be invaluable for any researchers planning to establish new prospective studies, or indeed those planning to enhance existing ones, by providing clear and practical advice on how to deliver on their aims.

**—Rory Collins FRS, University of Oxford, UK**

Epidemiology, including study design, conduct, data processing and analysis and interpretation, has been well described in numerous textbooks. In contrast, the book by Professor Chen clearly articulates areas rarely emphasized but of increasing importance in the interface with genomics and precision medicine.

Chen and colleagues’ considerable experience in prospective biobank studies has led to this seminal contribution. Their clear and comprehensive treatise includes case studies about challenges and innovative solutions in planning, establishing, and managing biobank studies.

This important and timely contribution will inform future studies in different settings by recognizing pitfalls of research using “real world” big data, often collected or generated without the similar scientific rigor well described in this book. For eager young students, and junior and senior faculty wishing to conduct research, this book represents one of the most informative texts available about

contemporary research methodology concerning biobank studies. Chen and colleagues merit kudos.

—**Charles H. Hennekens MD, Dr.PH, Florida Atlantic University, USA**

Prospective cohort studies are the cornerstone of epidemiology and play an indispensable role in identifying both genetic and non-genetic determinants of major chronic diseases. With advances in big data and “omics” technology, many large prospective cohort studies have been, or are currently being, established around the world, which should greatly advance the development of precision medicine.

Establishment of large prospective studies, especially those involving biological samples, is a major undertaking and requires careful planning and effective management. This book by Chen and colleagues provides a highly informative practical guide on design and delivery of a successful prospective study in diverse settings. Importantly, the book outlines in a succinct manner many of the challenges and practical solutions to some of the major issues involved in such studies.

This book not only provides practical advice, but also describes scientific considerations that underpin such approaches. It should also serve as a valuable reference for students and health researchers to enhance their understanding of advances in epidemiology.

—**Liming Li MD, MPH, Peking University, China**

Most developing countries have death rates in middle age in large excess of those in Western non-smoking populations. Much remains to be uncovered about the avoidable causes of chronic diseases such as vascular, respiratory, and neoplastic disease that account for most of the adult mortality worldwide. This requires attention to better quantification in prospective studies of “established” hazards, from tobacco, blood pressure, blood lipids, and adiposity, but also to discovery of new (mostly blood-based) genetic and biological factors.

Design and conduct of large prospective studies with biological samples require unique approaches. This deeply informative book by Professor Chen and his colleagues, based on their substantial and successful experience in China, will inform and enable similar studies elsewhere. The Indian Study of the Health of Adults is one such example, and hopefully prospective blood-based studies will begin in Africa. Studies in these populations can examine chronic infections (such as tuberculosis and malaria) adding to worldwide evidence about the avoidability of mortality in middle age.

—**Prabhat Jha MBBS, DPhil, University of Toronto, Canada**

—**Rajesh Dikshit MD, PhD, Tata Memorial Centre, Mumbai, India**

Prospective cohort studies are the most important source of observational evidence to identify common causes of major chronic diseases of adults. Attention to detail in the study design, implementation, and data management is critical to the success of such studies, particularly for large cohorts involving collection of biological samples.

This book by Chen and colleagues, to my knowledge, is the first reference book that describes details of some key practical components that investigators need to know and consider while planning and conducting prospective studies. In addition to practical advice, the book also provides the scientific basis for important aspects of study design and methodology, including the need for “periodic resurveys” to take account of “regression dilution bias” and the addition of study enhancements.

As a long-term collaborator of Prof. Chen and his colleagues at the University of Oxford, I believe that this “Practical Guide,” which is chiefly based on studies in China, will guide both on-going and future prospective studies worldwide.

—**Junshi Chen MD, China National Centre for Food Safety Risk Assessment, China**

This informative, practical handbook brings prospective epidemiologic research into the twenty-first century by focusing on that most recent of developments, the population “biobank.” Facilitated by—and indeed themselves stimulating—extraordinary advances in high-speed computing, data standardization and linkage, and robotic biospecimen repositories, biobanks allow pursuit of epidemiologic questions at a previously unimaginable size and scale.

Conducting such large studies involves operating at an almost “industrial” approach that may at first seem inimical to the scientific method, but is in fact supportive of much more ambitious goals than were feasible with traditional cohort studies. Standardization and harmonization of measures, inclusion of a wide range of sociodemographic groups and exposures, and innovative yet secure methods for data sharing all permit rapid, facile analyses to produce novel insights of truly practical value. That Chen et al. have captured the motivations and methods of such studies in such a slim volume is an impressive achievement, and one well worth exploring.

—**Teri Manolio MD, PhD, National Human Genome Research Institute, NIH, USA**

Throughout the twentieth century observational epidemiology, particularly cohort studies, contributed meaningfully to our understanding of disease etiology. At the beginning of the twenty-first century revolutions in computing power, availability of scalable genetic and other omic data, and the accumulation of electronic health data allowed for the emergence of the mega-cohorts or biobank studies.

Mega-cohorts have been established in diverse populations, each combining centrally stored biospecimens with health and lifestyle data. Most cohorts have consented participants in a way that permits recontact, enabling the cohorts to be enhanced over time. This trend is continuing and newer mega-cohorts are being established around the world.

The book by Chen skilfully gathers valuable insights into the inner workings of biobanks in three ways. First, it provides a compendium of lessons learned as new mega-cohorts are being established. Second, it provides best practices for optimal use of these massive resources. Finally, it serves as a road map showing how these

cohorts with varied methods can interact. This interaction will be essential to make the most of these valuable resources for decades to come.

—**J Michael Gaziano MD, MPH, Harvard Medical School, USA**

Large-scale population biobanks have emerged as a central component of the twenty-first century biomedical research. Of these, the China Kadoorie Biobank established by Professor Chen and colleagues in Oxford and China is an international leader. Documentation of the rationale, methods, handling of biosamples and data for such a resource has not appeared in a complete yet readable format until now. This book will become a landmark in the way the first monographs from the Framingham and Seven Countries studies have been foundational texts in epidemiology for half a century. As a decreasing proportion of epidemiologists engage in actual data collection—whilst being enthusiastic users of such data—this documentation is the nearest they will come to understanding the processes through which data are generated, which is essential for critical and scientifically valid approaches to their use. The world needs a network of population-based Biobanks, involving all age groups, and based in disparate environments. In this way, we can move to more robust inference regarding the determinants of individual and population health. By letting hundreds of Biobanks bloom such a mission can be advanced, and this book is a major contribution to helping that happen.

—**George Davey Smith FRS, University of Bristol, UK**

# Contents

<b>1 Population-Based Health Studies: An Overview of Principles and Methods . . . . .</b>	<b>1</b>
Derrick Bennett and Robert Clarke	
<b>2 Design, Implementation, and Management of Biobank Studies . . . . .</b>	<b>27</b>
Zhengming Chen	
<b>3 Planning, Organisation, and Management of Fieldwork in Biobank Studies . . . . .</b>	<b>51</b>
Ka Hung Chan, Kin Bong Hubert Lam, and Huaidong Du	
<b>4 Collection, Processing, and Management of Biological Samples in Biobank Studies . . . . .</b>	<b>77</b>
Iona Y. Millwood and Robin G. Walters	
<b>5 Monitoring Long-Term Health Outcomes of Biobank Participants by Record Linkages . . . . .</b>	<b>99</b>
Ling Yang and Zhengming Chen	
<b>6 Verification and Adjudication of Health Outcomes in Prospective Cohort Studies . . . . .</b>	<b>123</b>
Yiping Chen and Robert Clarke	
<b>7 Development and Application of IT Systems in Biobank Studies . . . . .</b>	<b>145</b>
Garry Lancaster, Simon Gilbert, and Xiaoming Yang	
<b>8 Management and Curation of Multi-Dimensional Data in Biobank Studies . . . . .</b>	<b>171</b>
Gary Sansome and Alex Hacker	

# Editor and Contributors

## About the Editor

**Zhengming Chen** is a Professor of Epidemiology at the Nuffield Department of Population Health, University of Oxford, UK. He qualified in Medicine at the Shanghai Medical University in China in 1983 and obtained a DPhil in Epidemiology from the University of Oxford in 1992. His main research focus has been on the determinants of chronic disease and advancement of evidence-based medicine. Since the mid-1990s, he has initiated and conducted several large prospective cohort studies and randomised trials of treatment for heart attack, stroke and cancer, yielding important findings that have since changed clinical practice and health policies worldwide. Most notably, he initiated, designed and co-directed, as the UK study principal investigator, the China Kadoorie Biobank's (CKB) study of over 500,000 adults, which has applied innovative approaches and methodologies for participant recruitment, data collection, quality assurance, study management and follow-up. He currently leads a large research team in Oxford, with expertise in epidemiology, population health, genomic medicine and data science.

## Contributors

**Derrick Bennett** MRC Population Health Research Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK

**Ka Hung Chan** Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK

**Yiping Chen** MRC Population Health Research Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK

**Zhengming Chen** Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK

**Robert Clarke** Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK

**Huaidong Du** MRC Population Health Research Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK

**Simon Gilbert** MRC Population Health Research Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK

**Alex Hacker** Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK

**Kin Bong Hubert Lam** Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK

**Garry Lancaster** Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK

**Iona Y. Millwood** MRC Population Health Research Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK

**Gary Sansome** MRC Population Health Research Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK

**Robin G. Walters** MRC Population, Health Research Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK

**Ling Yang** MRC Population Health Research Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK

**Xiaoming Yang** MRC Population Health Research Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK



# Chapter 1

## Population-Based Health Studies: An Overview of Principles and Methods



Derrick Bennett and Robert Clarke

### Contents

1.1 Introduction .....	2
1.2 Observational Epidemiological Studies .....	4
1.3 Bias and Confounding in Observational Epidemiology .....	9
1.4 Experimental Studies .....	15
1.5 Approaches for Assessing Causality .....	17
1.6 Modern Developments in Prospective Biobanks .....	21
1.7 Accurate Reporting of Epidemiological Studies .....	22
1.8 Summary .....	23
References .....	24

**Abstract** Chronic non-communicable diseases (NCDs) are the major causes of premature death and disability in both high-income and low- and middle-income countries (LMICs). However, there is substantial variation in the age and sex-specific rates of major NCDs that are not fully explained by differences in the distributions of established risk factors, suggesting that other important causes remain to be discovered. Population-based epidemiological studies are needed for reliable assessment of lifestyle, biochemical and genetic determinants for NCDs, and for assessing prognosis and clustering of NCDs. Moreover, analysis of genetic variants for particular traits can be used to elucidate the causal relevance of particular exposures with disease and to anticipate the likely effects of treatments. Epidemiological studies conducted in diverse populations with prolonged follow-up for both fatal and non-fatal disease outcomes can provide important evidence about the causes of NCDs that may inform disease prevention strategies globally. The aim of this chapter is to provide readers with an overview of basic concepts and the epidemiological principles that underlie the design and conduct of epidemiological studies, including the chief strengths and limitations of different study designs. Moreover, it will highlight the importance of large prospective biobank studies,

---

D. Bennett · R. Clarke (✉)

Big Data Institute Building, Nuffield Department of Population Health, Old Road Campus,  
University of Oxford, Oxford, UK

e-mail: [robert.clarke@ndph.ox.ac.uk](mailto:robert.clarke@ndph.ox.ac.uk)

© Springer Nature Singapore Pte Ltd. 2020

Z. Chen (ed.), *Population Biobank Studies: A Practical Guide*,

[https://doi.org/10.1007/978-981-15-7666-9\\_1](https://doi.org/10.1007/978-981-15-7666-9_1)

which involve assessment of a sufficiently large number of participants, together with strict control of bias and confounding to be able to detect moderate relative risks of major disease outcomes reliably.

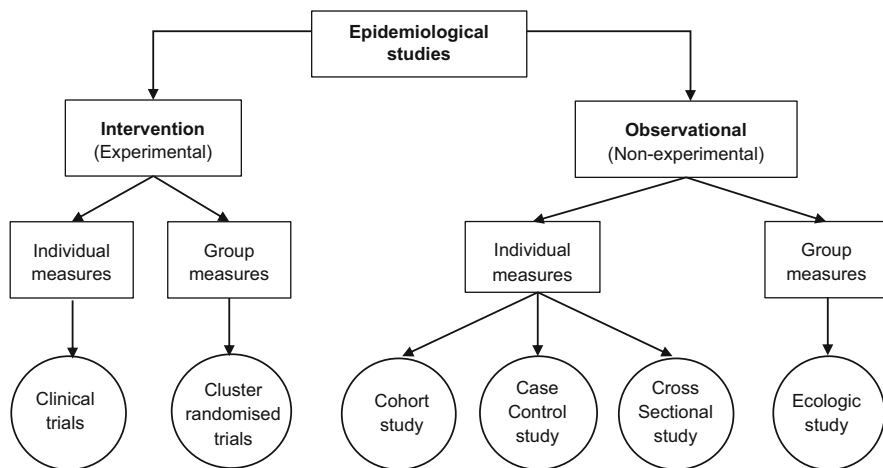
**Keywords** Epidemiological studies · Cohort studies · Case-control studies · Cross-sectional studies · Ecological surveys · Randomized trials · Biobanks · Causality

## Abbreviations

CHD	Coronary heart disease
CKB	China Kadoorie Biobank
DAG	Directed acyclic graphs
DNA	Deoxyribonucleic acid
GWAS	Genome-wide association studies
IHD	Ischaemic heart disease
IV	Instrumental variable
K	1000
MR	Mendelian randomization
MVP	Million Veterans Project
NCD	Non-communicable disease
RCT	Randomized controlled trials
STROBE	Strengthening the Reporting of Observational studies in Epidemiology
UK	United Kingdom
UKB	UK Biobank
USA	United States of America

## 1.1 Introduction

Epidemiology is the study of the occurrence and distribution of diseases or health-related events in human populations and the application of such knowledge for disease prevention and improvement in health. Epidemiological studies typically involve the assessment of the frequency, determinants and consequences of diseases in populations. Estimating disease frequency requires defining criteria for diagnosis, reliable mechanisms for follow-up to ascertain incident cases of such diseases in defined populations. Distribution refers to analyses by person, place and time, while determinants are sometimes referred to as exposures, risk factors or risk markers depending on their causal relevance to disease. Chronic non-communicable diseases (NCDs) are the major causes of death and disability worldwide, and the majority of NCDs worldwide now occur in low- and middle-income countries (Kyu et al. 2018). The chief objectives of epidemiology are to: (1) study the natural history of both communicable and non-communicable diseases; (2) determine the frequency of individual diseases; (3) identify patterns or trends in disease occurrence; (4) establish



**Fig. 1.1** Main study designs used in population health research

the causes of diseases; and (5) evaluate the effectiveness of measures for prevention of major chronic diseases.

Epidemiological studies select the most appropriate study design and scientific methods to determine the causes and consequences of major diseases. Typically, scientific methods use observations and theories to formulate and test particular hypotheses. Associations of exposures with disease can be causal, but may also be artefacts of chance, bias (erroneous associations) or confounding (spurious associations). Hence, studies need to evaluate the reliability and validity of any relevant data collected on both exposures and disease outcomes. Moreover, before concluding that associations of individual exposures with particular diseases are causal, analyses should exclude the possible effects of chance, bias or confounding. It is important to adopt strategies to prevent bias when designing studies, and control for confounding before making inferences about causality. Different study designs are used to investigate the risks of disease associated with risk factors in different settings. In general there are two broad categories of designs in epidemiological research, namely observational (i.e., non-experimental) and intervention (i.e., experimental) studies. For each, there are specific types of study and the selection of the most appropriate study design (Fig. 1.1) depends critically on the research question being addressed. This chapter provides an overview of basic concepts and the epidemiological principles that underlie the design and conduct of population studies, along with modern developments and the scientific potential of large prospective biobank studies, which is the main focus of the book.

## 1.2 Observational Epidemiological Studies

Observational studies are conventionally classified into descriptive and analytic studies. The aims of descriptive epidemiological studies (i.e., ecological studies, cross-sectional studies) are to provide reliable estimates of the frequency of diseases in different study populations. Differences in disease frequency and risk exposures between populations or subgroups of the population (e.g., ethnic group or geographic location) may help to identify risk factors for such diseases. Analytic epidemiological studies typically examine the shape and strength of associations of putative risk factors with specific diseases. There are two major types of such studies (i.e., case–control studies and prospective cohort studies), but both types of study essentially compare the risks of disease in exposed versus unexposed groups (Table 1.1). This section provides a brief overview of the different types of study designs and illustrates the strength and limitations of such study design using examples from contemporary large studies.

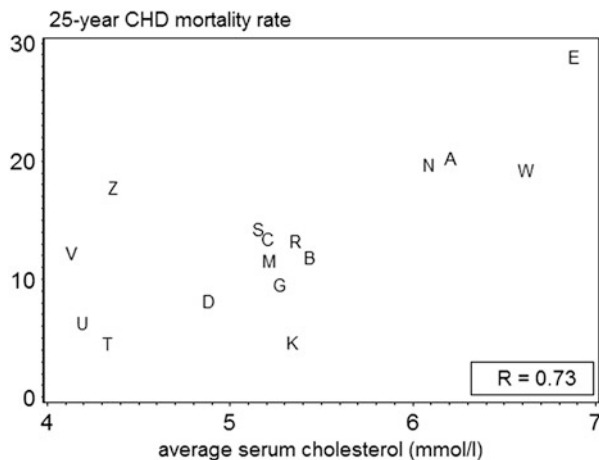
### 1.2.1 Ecological Studies

Ecological studies are typically designed to investigate the associations between exposures and disease outcomes in populations (or groups) rather than in individuals. Ecological studies identify correlations between exposures and diseases, or surrogate measures of diseases, by geographical region (countries, regions or cities within countries), or by time (calendar period or birth cohort). In particular, ecological studies examine correlations between average levels of exposures in different groups of people with the overall frequency of disease within such groups within or between populations (Hennekens and Buring 1987). Some risk factors for disease operate at a population level and under certain specific circumstances may cause

**Table 1.1** Glossary of key epidemiological terms

Terminology	Meaning
Prevalence	Proportion of individuals in a population who have a disease at a specific point in time or probability that an individual will be ill at a particular time point
Incidence	The number of new cases of disease that develop in a population of individuals during a specified time interval
Confounding	Underestimation or overestimation of the presence or strength of an association between an exposure and disease due to another factor
Bias	Systematic error that distorts an association between an exposure and disease in an epidemiological study
Effect modification (interaction)	Effect of a risk factor on disease varies according to the levels of another factor
Representativeness	The degree to which a sample of a population accurately reflects the characteristics of the total or targeted population

**Fig. 1.2** Total cholesterol and risk of coronary heart disease in Seven Countries study (Keys 1980)

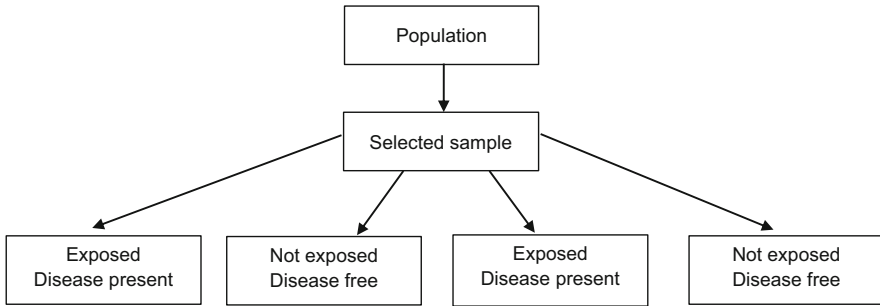


disease. However, risk factors are more likely to cause disease if they are also determinants of exposure to individual level risk factors for disease. Ecological studies play an important role in identifying relevant public health problems to be addressed, and in generating hypotheses to test the likely causes of such diseases. For example, the Seven Countries study (Keys 1980) first highlighted the strong association between blood levels of total cholesterol and risk of coronary heart disease (CHD), by plotting mean levels in 16 distinct populations from seven countries with risk of CHD. The results demonstrated a strong positive relationship between higher levels of total cholesterol and risk of CHD throughout the range studied (Fig. 1.2).

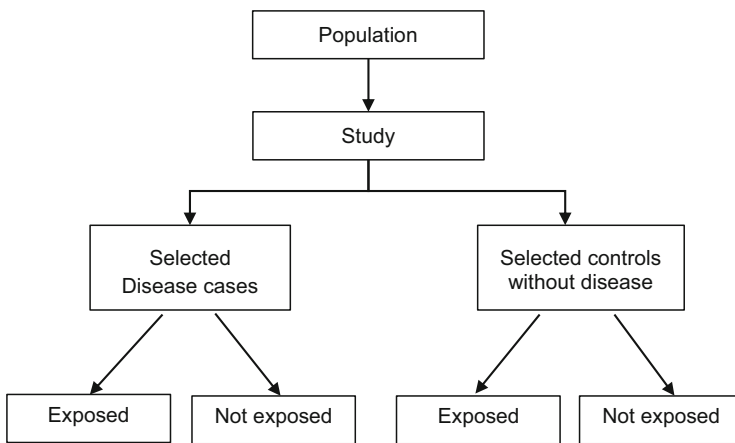
The Seven Countries study prompted the establishment of multiple prospective cohort studies in different populations that subsequently confirmed these associations at an individual level over a wide range of total cholesterol values. However, ecological studies have major limitations, including the *ecological fallacy* that occurs when associations observed in groups are contradicted at an individual level. For example, chronic hepatitis B viral (HBV) infection was found to be associated with higher levels of blood cholesterol in ecological studies in China, but when analysed at the individual level the converse was true, consistent with effects of chronic HBV infection on metabolism of cholesterol (Chen et al. 1993).

## 1.2.2 Cross-Sectional Surveys

Cross-sectional studies typically describe the health of a population at one time and, hence, such studies need to be based on representative samples of the underlying population in order to make valid inferences about frequency of risk exposures or diseases in general populations, or about associations of risk factors with disease (Fig. 1.3). Cross-sectional surveys assess the prevalence of disease and the



**Fig. 1.3** Cross-sectional study design



**Fig. 1.4** Case-control study design

prevalence of risk factors at a fixed point in time and provide a “snapshot” of both risk factors and diseases simultaneously in a defined population (Kirkwood and Sterne 2003). However, a major limitation of cross-sectional studies is that they lack the appropriate temporal sequence between exposures and diseases and, hence, can only reliably assess disease prevalence.

### 1.2.3 Case-Control Studies

A case-control study (also known as a retrospective study) involves a group of individuals with a disease (cases) and another group without the disease (controls) (Fig. 1.4). The frequency of past exposure to a putative risk factor is then determined in both groups (usually by asking the participants to report their lifestyle or other traits using either self-administered or interview-administered questionnaires). Moreover, they may also involve comparing differences in mean levels of clinical

measures (anthropometric traits) or blood levels of exposures (e.g., biochemical or genetic measures) between cases and controls. If more cases than controls are exposed, then this suggests that such exposures may be risk factors for the disease (Kirkwood and Sterne 2003).

The two aspects that are particularly important when designing case–control studies: (1) explicit diagnostic criteria for cases (including eligibility criteria used for selection of cases) and (2) controls should come from the same “source population” as the cases, and their selection should be independent of the exposure or exposures of interest (Kirkwood and Sterne 2003).

Case–control studies, although intuitively simple, can be challenging to design and require strict criteria for the selection of cases and controls (Hennekens and Buring 1987). After clearly defining cases and controls, the relevant data on exposures and other important variables have been collected, the investigator then estimates a measure of effect for the main exposure with risk of disease. The odds ratio and its associated confidence intervals are used as a measure of the strength of the association between an exposure and disease outcomes, where exposures may be continuous, dichotomous or quantiles of a distribution (e.g., quintiles) (Box 1.1).

#### **Box 1.1 Effect Size Metrics used in Analytic Epidemiology**

*Odds Ratio:* This represents the ratio of the odds of exposure in cases compared with the odds of the exposure in controls.

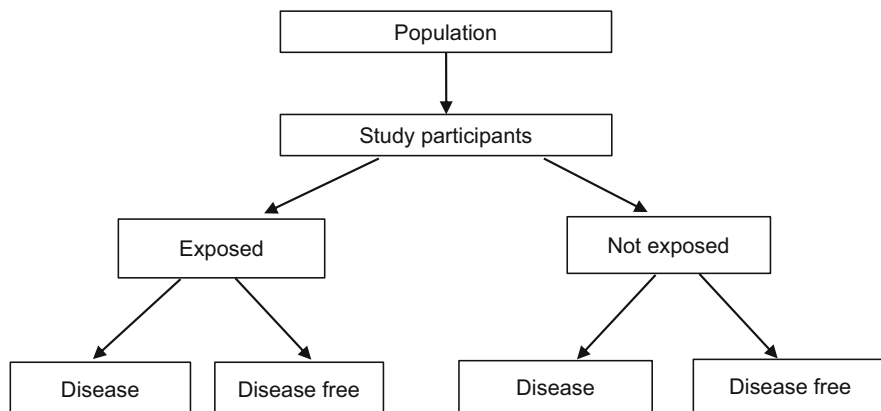
*Relative Risk or Risk Ratio:* The ratio of the risk in the exposed to the risk in the unexposed.

*Rate Ratio:* The ratio of incidence rate in the exposed to the incidence rate in the unexposed.

*Risk difference:* The risk difference is calculated by subtracting the cumulative incidence in the unexposed group from the cumulative incidence in the group with the exposure.

### **1.2.4 Prospective Cohort Studies**

Prospective cohort studies are characterized by selection of participants in which risk factors or exposures are recorded at enrolment before the onset of disease (Fig. 1.5). Hence, prospective studies have the appropriate temporal sequence between exposures and outcomes to assess causality whereby measurements of exposures always precede the onset of disease outcomes (and risk factors are recorded before the onset of disease) (Hennekens and Buring 1987). Prospective studies are superior to standard case–control studies in which exposures and disease are assessed concurrently and all participants should have an equal risk of developing any disease during follow-up. Moreover, associations can also be assessed with a wide range of diseases. Baseline data are collected on all participants including assessment of exposures that may alter the risk of developing the disease, including age, sex,



**Fig. 1.5** Cohort study design

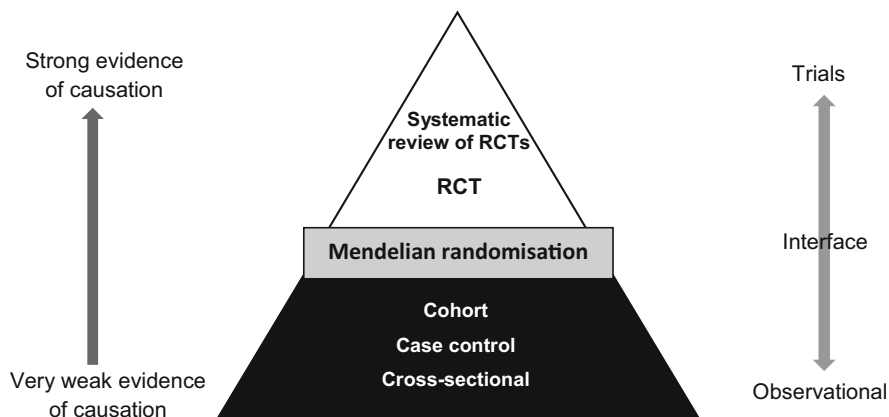
medical history and concomitant use of medication that may alter levels of exposures. The key requirement for contemporary cohort studies is to study a sufficiently large number of individuals (e.g., >100,000 participants) with a wide range of exposures to enable investigators to detect moderate relative risks for major diseases. However, reliable detection of weak associations, saying with relative risks <twofold, requires strict control of chance, bias or confounding, which underpins the design, conduct and data analysis of contemporary cohort studies.

One of the major challenges of conducting prospective cohort studies is their high cost, as prospective studies typically require a large number of participants to be followed up for a prolonged period of time. Consequently, many prospective studies have a high risk of *loss to follow-up*, whereby investigators lose contact with some of the study participants. It is important to minimize *loss to follow-up* to reduce the risk of missing data. When *loss to follow-up* of many individuals occurs, the internal validity of such studies is attenuated, as there may be systematic differences in disease rates or risk factors for those who are lost to follow-up versus those who remain in the study. Strategies to minimize loss to follow-up include automated processes (i.e., so-called passive follow-up) to enable subsequent tracking and maintaining periodic contact with study participants (Hennekens and Buring 1987). Another major challenge is to obtain reliable information about diagnoses of different health outcomes. In contrast with case-control studies, it may not always be feasible or possible for prospective studies to collect health outcomes directly from hospitals.

### **1.2.5 Nested Case-Control Studies**

For rare diseases, the odds ratio calculated from a case-control study provides a reliable estimate of the relative risk (rate-ratio) as defined in a cohort study





**Fig. 1.6** Hierarchy of evidence of different study designs

(Hennekens and Buring 1987). However, if the case–control study is nested within a cohort study (i.e., samples of cases and controls are selected from a well-defined cohort or prospective study, with appropriate matching for key variables), the resulting odds ratios will approximate the rate ratios for the “source population” (Hennekens and Buring 1987). *Nested case–control* studies are typically used to investigate the associations of biological factors with specific diseases, which involve assays of selected stored samples in a random sample of selected cases and controls rather than in the overall cohort. Such an approach is more efficient in the setting of constraints of funding or limited relevance of such biomarkers for other diseases in the cohort. To facilitate appropriate future use of such data in controls for individual diseases, the controls should be randomly selected from the cohort, and such a study design is referred to as *case-cohort* study.

Overall, the evidence from different types of study design addressing a particular hypothesis has different relevance for causal inference, which can be ordered in a hierarchy (Fig. 1.6), with ecological studies at the bottom and trials (RCTs), particularly meta-analysis of all such RCTs, at the top.

### 1.3 Bias and Confounding in Observational Epidemiology

Observational epidemiological studies are susceptible to bias and may produce results that differ systematically from the truth. Bias is typically classified by sources as either selection bias or information bias. Table 1.2 summarizes the main types of bias and examples of each type of bias together with conventional strategies to minimize such biases at various stages of studies. Case–control studies are more susceptible to bias than prospective studies. It is prudent to use incident (new cases) rather than prevalent cases (i.e., existing cases) in case–control studies to minimize

**Table 1.2** Major types of bias and strategies for prevention of bias

Type	Examples	Prevention strategies
Information bias	Observer bias	Masking of participants from identity of exposures
	Interviewer bias	Concealment of identity of exposures from interviewers
	Recall bias	Collection of data from medical records
	Reporting bias	Masking or blinding of participants
	Performance bias	Masking or blinding of assessments
	Detection bias	Masking or blinding of assessments
	Sampling bias	Avoid volunteers/use rigorous inclusion criteria
Selection bias	Reverse causality	Exclude events occurring in the initial years after baseline
	Allocation bias	Clear inclusion criteria/proper randomisation
	Loss to follow-up	Use tracing methods to minimize loss to follow-up

risks of prevalent cases changing their behaviour or the onset of disease altering the levels of exposures of interest (referred to as reverse causality bias: Table 1.2).

### 1.3.1 Confounding

Confounding is a limitation of most observational epidemiological studies, which may result in spurious associations or may underestimate or overestimate the strength of associations due to the presence of some other correlated variables. In descriptive epidemiological studies, death rates and disease rates are usually strongly related to age and sex. Hence, comparisons of disease rates critically depend on the age and sex-specific composition of the underlying populations being compared. Therefore, it is potentially misleading to use overall crude disease rates when comparing different populations (unless they have an identical age and sex composition). Standardization is a method (Box 1.2) used in descriptive epidemiology to enable reliable comparisons of measures between populations or subgroups after adjustment for differences in the age and sex structure of the populations or subgroups being compared (Hennekens and Buring 1987).

#### **Box 1.2 Controlling for confounding in descriptive epidemiological studies**

The comparison of crude mortality or morbidity rates is often misleading because the populations being compared may differ significantly with respect to certain underlying characteristics, such as age or sex that will affect the overall rate of morbidity or mortality.

One method of overcoming the effects of confounding variables such as age is to simply present and compare the age-specific rates. While this allows

(continued)

**Box 1.2** (continued)

for a more comprehensive comparison of mortality or morbidity rates between two or more populations, as the number of stratum specific rates being compared increases, the volume of data being examined may become unmanageable.

It is, therefore, more useful to combine category specific rates into a single summary rate that has been adjusted to take into account its age structure or other confounding factor. This is achieved by using the method of standardization.

There are two methods of standardization commonly used in epidemiological studies, and these are characterized by whether the standard used is a population distribution (*direct method*) or a set of specific rates (*indirect method*). Both direct and indirect standardization involve the calculation of number of expected events (e.g., deaths), which are compared to the number of observed events.

In analytical epidemiology, investigators may control for confounding either at the design stage or in the analysis stage using statistical techniques such as stratification or multivariable regression analyses (Kirkwood and Sterne 2003). Matching controls to cases at the design stage of a case-control study is one approach used to minimize the effects of confounding. Matching aims to minimize confounding by reducing systematic differences between those with the disease and those without the disease in terms of the confounding factors (Kirkwood and Sterne 2003). Matching should be used carefully as cases and controls should not be matched on variables that the investigator may be interested in assessing possible associations with the disease outcomes. Not only should key exposure variables not be used to match cases and controls, but closely related variables should also not be used.

When examining the effects of confounding factors, analyses should assess the magnitude of the effect size (e.g., odds ratio or rate ratio) before and after adjusting for differences in levels of potential confounding factors using multivariable regression models (Kirkwood and Sterne 2003). If the magnitude of the adjusted effect size is largely unaltered after adjustment for known confounding factors (e.g., age, sex, smoking, education), the likelihood of confounding is low and investigators may conclude that the results are reliable. However, one cannot fully exclude residual confounding due to some unknown or unmeasured confounders, or to incomplete adjustment for known confounders.

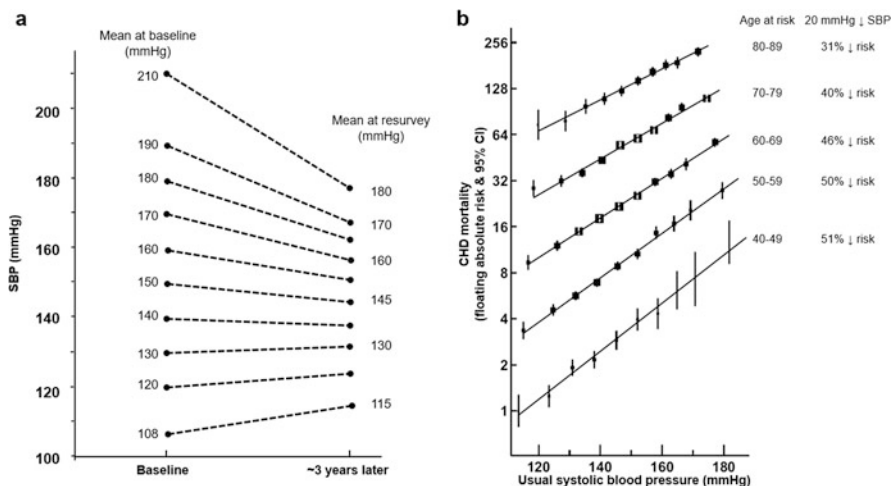
### 1.3.2 *Reverse Causality Bias*

Reverse causality bias occurs when onset of disease affects the levels of exposures in observational studies (e.g., reduced weight due to pre-clinical cancer), leading to

spurious associations (e.g., inverse association of adiposity with cancer risk). Although reverse causality bias is a common problem in retrospective case–control studies, it may still represent a major challenge in prospective studies. Strategies to quantify the magnitude of reverse causality bias include comparison of associations of exposures with disease in all study participants with available data with subsets after excluding disease events occurring within the first 5 years of baseline survey or stratifying by those with or without concomitant diseases that may alter levels of exposure. For diseases with a longer latent periods, such as dementia or Parkinson’s disease, it may be necessary to exclude events occurring during the first 10 years (or longer) of follow-up in order to avoid the full effects of reverse causality bias (Floud et al. 2020). In addition to excluding individuals with prior disease, prospective studies should also consider excluding individuals with self-reported poor health to fully exclude the effects of reverse causality. For example, in the UK Million Women Study, unhappiness was strongly associated with increased risk of death from IHD or cancer after rigorous adjustment for all relevant confounding factors. However, after excluding individuals with self-reported poor health at baseline, which was related to unhappiness, the apparent associations of unhappiness with excess risks of death completely disappeared (Liu et al. 2016).

### ***1.3.3 Regression Dilution Bias***

Analysis of observational studies assumes that both exposures and disease outcomes are measured without error or variation. However, in practice, measurement error and biological variation are common. Purely random errors in exposures will systematically underestimate the strength of associations of such exposures with disease outcomes, referred to as “regression dilution bias” (MacMahon et al. 1990). Regression dilution bias is a phenomenon whereby measurements at the extreme of a distribution at a single measurement are closer to the mean on repeat measurements. Regression dilution bias is common and may reflect measurement error, biological variation, effects of subclinical disease or treatment. Correction for regression dilution bias requires repeat measurements in a random sample of the population collected at a later period of follow-up. Failure to correct for regression dilution bias will systematically underestimate the strength of associations between exposures and disease outcomes. Figure 1.7 illustrates the importance of regression dilution bias when assessing the associations of systolic blood pressure (SBP) with risk of death from IHD in the Prospective Studies Collaboration (PSC) meta-analysis of prospective studies. When assessing the shape of associations of SBP, investigators typically classify individuals into 10 equal sized groups and relate these on the x-axis to risk of IHD or stroke on the y-axis. Figure 1.7 (panel [a]) shows that mean levels of baseline SBP varied from 108 to 210 mmHg. However, when individuals in these baseline-defined groups were re-measured after an interval of 3 years, the mean difference between the 10 groups now only varied from 115 to 180 mmHg between the extreme groups. By plotting the disease risks against the baseline-defined groups would have



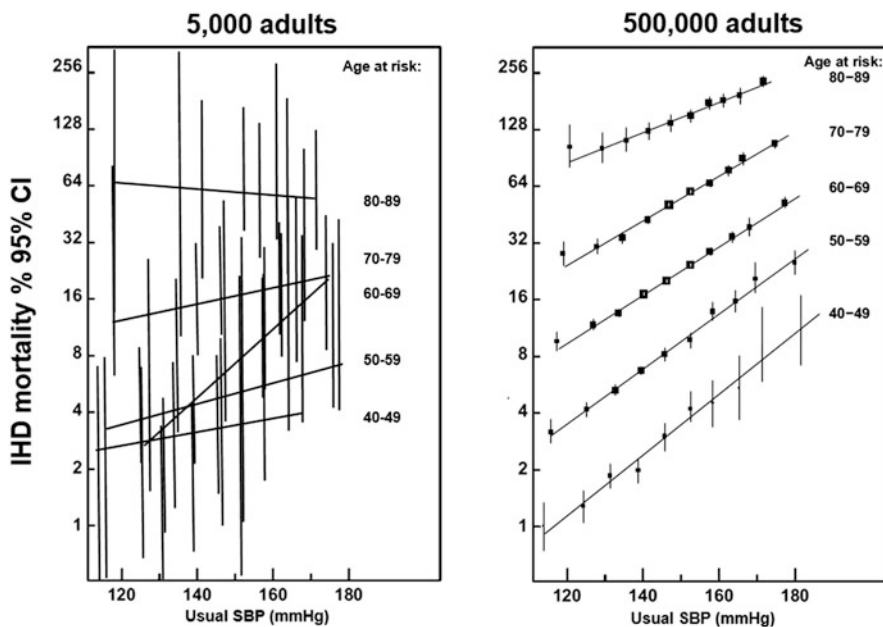
**Fig. 1.7** SBP and risk of IHD mortality. Panel (a) shows mean levels of deciles of SBP at baseline and at resurvey at 3 years after baseline in baseline-defined groups. Panel (b) shows the associations of usual SBP levels with risk of IHD mortality in age-specific groups (Prospective Studies Collaboration 2002)

substantially underestimated the strength of the association of SBP with risk of IHD by 50%. In the PSC meta-analysis (Fig. 1.7 Panel [b]), the associations of SBP with risk of IHD were plotted against *usual* levels of SBP (i.e., mean levels at resurvey of baseline-defined groups) and the associations were linear throughout the range studied and the strength of such associations was twofold stronger than those plotted against baseline levels of SBP.

In addition to measurement error, regression dilution bias may reflect other causes of within-person variability including ageing, onset of subclinical disease, or initiation of treatment. Prolonged follow-up of middle-aged cohorts is particularly susceptible to time-dependent regression dilution bias whereby the phenomenon becomes even more extreme with longer intervals between measurements and incident disease outcomes (Fig. 1.7, Panel [b]) (Clarke et al. 1999). Application of time-dependent correction for regression dilution bias in the PSC in each decade of risk in ~one million adults from 61 prospective studies demonstrated a twofold difference in death rates from stroke, IHD and other vascular causes at ages 40–69 years, but death rates at 80–89 years were only half as extreme as those at ages 40–69 years. Importantly, the associations of SBP with mortality were linear throughout the range studied with no evidence of a threshold down to at least 115/75 mmHg (Prospective Studies Collaboration 2002). These results of this meta-analysis highlighted the importance of large prospective studies with periodic repeat measurements of exposures in a random subset of participants during follow-up to correct for time-dependent regression dilution bias (Prospective Studies Collaboration 2002).

### 1.3.4 Need for Large Studies

The play of chance is frequently overlooked in observational epidemiology despite use (or inappropriate use) of  $p$ -values and 95% confidence intervals around effect sizes of associations of exposures with disease outcomes. The key requirements to minimize the effects of chance are to maximize sample size, minimize the use of multiple testing and seek replication of any completely novel associations in independent populations wherever possible. Figure 1.8 shows the associations of usual SBP with risk of IHD by age at risk in the PSC meta-analysis of prospective studies in 5000 (the size of Framingham study) versus 500,000 (the size of China Kadoorie Biobank) adults who were randomly selected from the PSC dataset. The results prompted the need for very large prospective studies to be able to reliably assess the shape and strength of the associations of blood pressure with IHD in age-specific groups. When investigating a more modest association of certain exposures (e.g., genetic factors) with common diseases, the requirement for large sample size will become even more critical.



**Fig. 1.8** Age-specific associations of SBP with IHD mortality in 5000 vs 500,000 adults in the Prospective Studies Collaboration (Prospective Studies Collaboration 2002)

### 1.4 Experimental Studies

The major advantage of experimental studies or trials is that they are less susceptible to confounding, because the investigator determines, usually through computer program, who is exposed and who is unexposed.

In trials, the exposures are randomly allocated at enrolment and provided that the number of individuals randomized is sufficiently large then both measured and unmeasured confounding should be equally balanced between the allocated groups (Fig. 1.9). Parallel group designs are the most widely used design typically to test the efficacy and safety of novel drug treatments for specific diseases, whereby some participants with disease are randomly allocated to take active treatment and others to placebo (or usual care) and both groups are followed up to record disease outcomes. The two groups are then compared prospectively for incidence of disease and other important outcomes (“endpoints”) of interest. Some of the key metrics that can be derived from RCTs are described in Box 1.3.

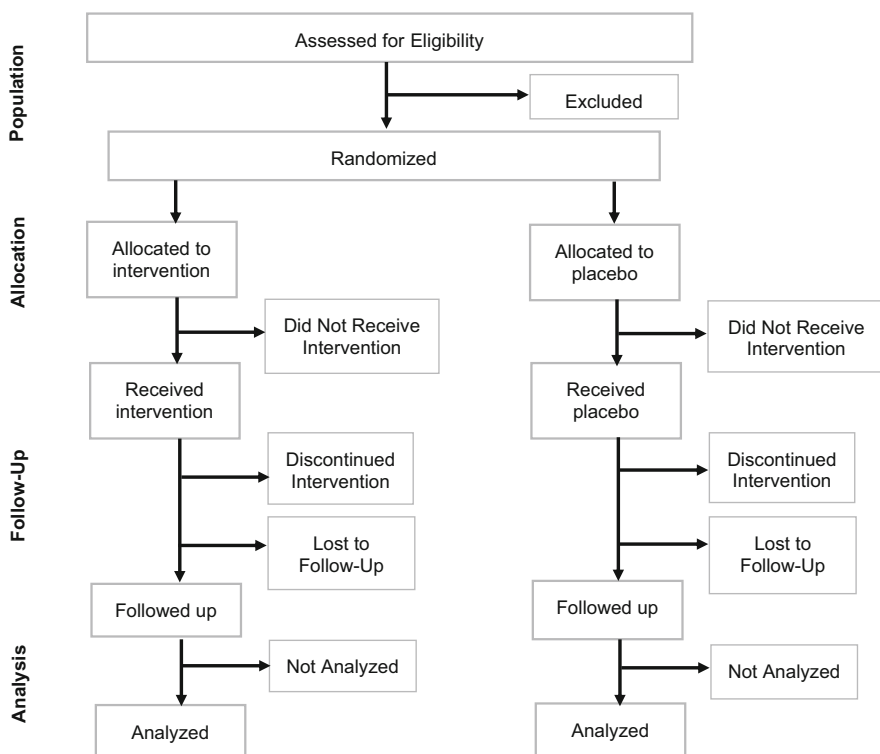


Fig. 1.9 Experimental studies

**Table 1.3** Key properties of randomized controlled trials

Requirement	Rationale for the requirement
Proper randomization	Allocation to the particular intervention should be done by a probabilistic random process that ensures balance for known and unknown factors and should be non-discoverable
Minimize bias	Patients, clinicians and outcome evaluators should be unaware of the allocation to a particular intervention (blinding or masking), and an adequate comparator group should be used. This is necessary so that there is no risk of knowledge of the treatment received, as this can influence the assessment of the patient
Intention to treat analyses	The treatment efficacy is based on treatment allocated rather than treatment received—That is, the statistical analyses are performed according to the randomized groups regardless of adherence and prevents bias
Maximize sample size and minimize the number of data-driven subgroup analyses	Essentially the more the data is divided up, the more false-positive results are likely to be observed due to chance. It is important to pre-specify intended analyses as analysing many based on data-driven subgroups can produce chance results

**Box 1.3 Metrics used in Randomized Controlled Trials**

*Relative Risk or Risk Ratio (RR)*: The ratio of the risk of a given event in treated group of participants compared to the control group.

*Relative Risk Reduction (RRR)*: The proportion of the initial or baseline risk which was eliminated by a given treatment/intervention or by avoidance of exposure to a harmful factor. The RRR is calculated as  $(1 - RR) \times 100\%$ .

*Odds Ratio (OR)*: The ratio of the odds of a given event in the treated group of participants compared to the control group.

*Absolute risk reduction (ARR)*: The difference in risk of a given event, between two—the treated and control groups.

*Number needed to treat (NNT)*: The number of patients needed to be treated in order to prevent one additional event (and is estimated by  $1/ARR$ ).

In general, high quality RCTs include: (1) proper randomization; (2) appropriate controls and blinding (masking) of participants; (3) intention-to-treat analyses; (4) pre-specification of a limited number of relevant subgroups to investigate effect modification; and (5) well-powered to minimize any chance effects (Baigent et al. 2020). Further details and rationale for the key design issues in RCTs are provided in Table 1.3. In addition to parallel group designs, other trial designs (factorial trial, cross-over trials or cluster randomized trials) can also be informative and selection of



the appropriate design depends on the research question, type of treatment and disease outcomes being assessed (Kirkwood and Sterne 2003).

## 1.5 Approaches for Assessing Causality

Causality (also referred as causation) is the process by which an exposure (determinant or risk factor) contributes to incidence of a disease outcome, or increases the probability of developing, or dying from, a particular disease. The common diseases of adults (e.g., cancer, IHD, stroke) often have multiple causes, and with a few exceptions (e.g., smoking, blood pressure) each tends to have modest or weak effects on disease incidence. Before concluding that any association is causal, studies should confirm that association of exposures with disease or outcomes is valid and not due to chance, bias or confounding.

Both observational epidemiological studies and randomized controlled trials play an important role in assessing causal inference in medical research. In principle, observational epidemiological studies can only identify associations, but randomized controlled trials are considered to be the most reliable study design when assessing causality in medical research due to their prospective nature and rigorous control of both measured and unmeasured confounding factors (Collins and MacMahon 2001). There have been many promising associations derived from observational studies that have been completely refuted by randomized controlled trials (Lawlor et al. 2004). This lack of confirmation has been attributed to inadequate control for bias and confounding in observational epidemiological studies and prompted the need for more robust methods to assess the causal relevance of exposures with disease in population studies.

For many lifestyle or environmental risk factors (e.g., smoking, alcohol drinking), however, it is not possible or feasible to conduct randomized controlled trials in order to assess their casual effects on certain specific diseases. As such, only evidence from observational studies can be considered, guided by Bradford-Hill proposed criteria for a higher likelihood of causality, which has been widely used by epidemiologists when making inferences about causality (Table 1.4). Increasingly,

**Table 1.4** Bradford-Hill criteria for causation

Criteria	Probability of causation
Strength of association	Strong associations after appropriate adjustment for confounders
Consistency	Replication in different settings
Specificity	Related uniquely to a particular exposure
Temporality	Cause precedes the effect on disease
Biological gradient	Dose response relationship with disease
Coherence	Logical or reasonable findings
Experimental	Confirmation of association in randomized trials
Analogy	Similar findings in different settings

genetics analyses play an important role in helping to assess the likely cause–effect associations of particular exposures with particular disease outcomes.

### ***1.5.1 Causal Inference in Epidemiology***

Recognition of the importance of causal inference has a major influence on the design, conduct, analysis, and interpretation of observational epidemiological studies in recent years. Advances in methodology for assessing causality include a combination of counterfactual and probabilistic approaches to causation (Vandenbroucke et al. 2016). The resulting toolkit, including the use of counterfactual concepts and directed acyclic graphs (DAGs), has enabled some problems which were previously considered intractable to be addressed more reliably. Instrumental variable (IV) analyses utilize variables that robustly relate to exposures of interest in a way that are analogous to randomization to an exposure or control. Such variables, like any technique used for proper randomization, should not be related directly to the outcome or to potential confounders (i.e., other risk factors for the outcome). A particular type of IV analysis that uses genetic variants as a proxy for exposures, known as “Mendelian Randomization”, has become more widely used in epidemiological studies in recent decades.

### ***1.5.2 Mendelian Randomization***

The underlying principle of “Mendelian Randomization” (MR) is that use of genetic variants that either alter levels or imitate the biological effects of modifiable biomarkers that are causal for disease, then such genetic variants are likely to be associated with risk of that disease proportional to their effects predicted by the effects of the genetic variants on levels of the biomarkers (Smith and Ebrahim 2003). Details of the MR approaches, including the underlying assumptions are provided in the DAG in Fig. 1.10.

“Mendelian Randomization” refers to the random assortment of alleles during meiosis where deoxyribonucleic acid (DNA) is transferred from parents to offspring at the time of gamete formation, following a process defined by Mendel’s second law (Smith and Ebrahim 2003). Mendel’s second law assumed that the inheritance of any particular genetic variant in an individual’s DNA should be independent of other characteristics, thus, when individuals in the population are classified by genotypes, they should be similar in all respects with the exception of one group with genetically instrumented higher levels of biomarkers and another groups with genetically instrumented lower levels of biomarkers analogous to a randomized controlled trial (Fig. 1.11).

Publication of the results of large international consortia of genome-wide association studies (GWAS) that include data on millions of genetic variants in large

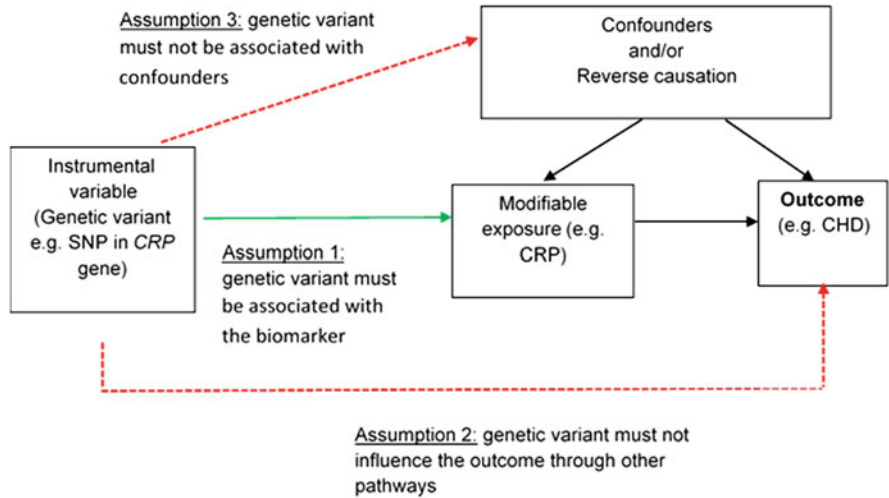


Fig. 1.10 DAG of the Mendelian randomization approach and key assumptions

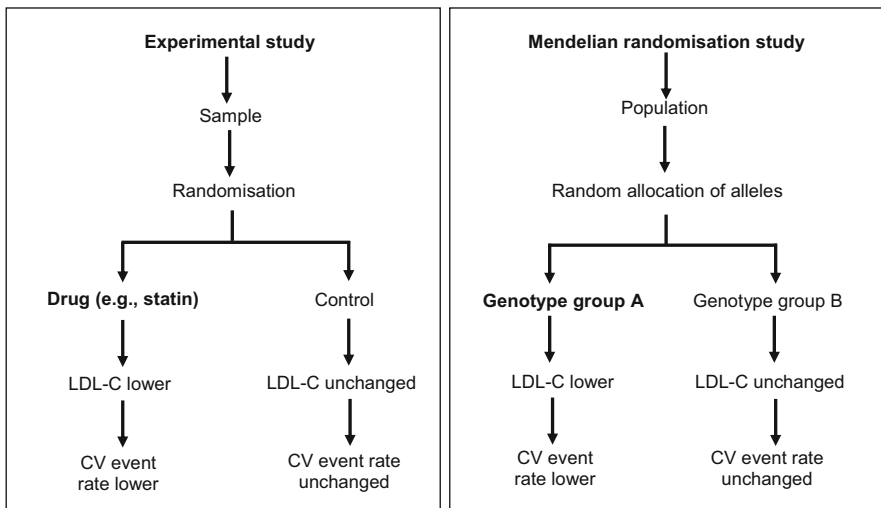
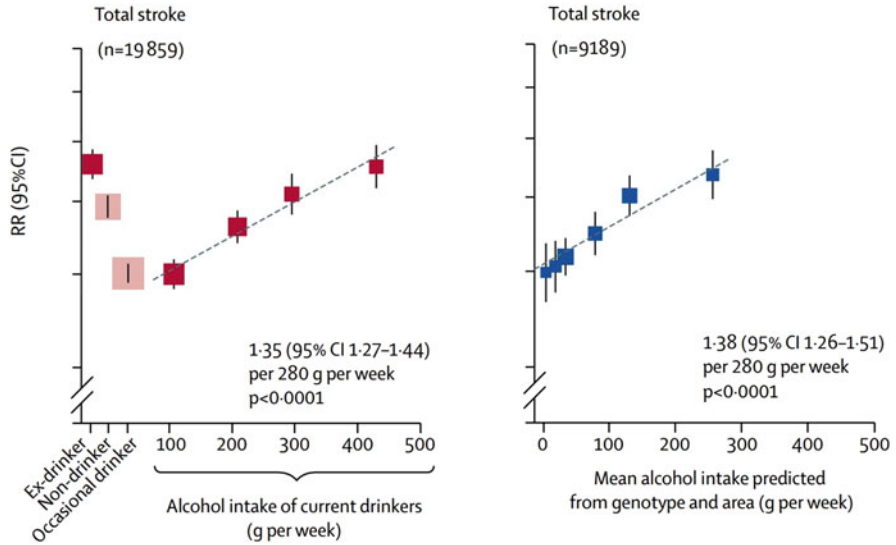


Fig. 1.11 Analogy between experimental studies and Mendelian randomization studies

numbers of individuals have enabled conduct of MR analyses to address the causal relevance of a wide range of exposures for multiple disease outcomes.

The CKB study used this MR approach to examine the causal relevance of moderate alcohol consumption for risk of stroke and heart attacks (Millwood et al. 2019). Previous studies had observed that people who drink one or two alcoholic drinks a day had previously been observed to have a slightly lower risk of developing a stroke and heart attack than non-drinkers, it was not known whether this was



**Fig. 1.12** Comparison of (a) observational versus (b) genetic associations of alcohol intake with risk of stroke in the China Kadoorie Biobank (Millwood et al. 2019)

because moderate drinking was “protective” for these disease outcomes, or whether it was because non-drinkers had other underlying health problems. In East Asian populations, there are common genetic variants that greatly reduce alcohol tolerability, because they cause an extremely unpleasant flushing reaction after drinking alcohol. Although these genetic variants greatly reduce the amount people drink, they are unrelated to other lifestyle factors such as smoking. Therefore, they can be used to study the causal effects of alcohol intake, analogous to a randomized trial of alcohol consumption. Figure 1.12 shows the application of MR analysis to assess the causal relevance of moderate alcohol consumption for risk of total stroke in the CKB study. In contrast with the typical J-shaped associations of self-reported alcohol consumption with total stroke in observational analyses, the genetic analyses using the predicted alcohol intake from genotypes showed approximately linear association with stroke risk throughout the range studied, thus reliably refuting any apparent protective effects of moderate drinking with such diseases. The discrepancy between the two results illustrates the challenges in the observational epidemiology for effective control for bias and confounding. Such MR analyses are particularly appropriate for testing hypotheses that are not possible to test in randomized trials (Millwood et al. 2019). Despite the proliferation of MR studies, the approach has strengths and limitations that have not been fully appreciated (Table 1.5).

**Table 1.5** Strengths and limitations of Mendelian randomization

Advantages	Disadvantages
Lack of confounding and reverse causality	Pleiotropy
Analogous to a randomized trial	Population stratification
Establishes potential causal relevance	Canalization/biological compensation
Provides additional biological insight	Linkage disequilibrium
Can assess the causal relevance of novel biomarkers and inform drug development	Lack of suitable genetic variants and potential weak instrument bias

## 1.6 Modern Developments in Prospective Biobanks

The advent of genomics has enabled population biobanks to play an increasingly important role in healthcare and the development of precision medicines (Ouellette and Tassé 2014). Biobanks are large prospective studies that have stored biological samples, together with follow-up for incident disease outcomes collected over prolonged periods. The combination of prospective cohorts with data (from questionnaires, interviews, clinical measurements and biological samples) on exposures collected prior to the development of disease and clinical data with information on incident disease outcomes provides valuable resources for assessing effects on multiple disease outcomes in relation to levels of single exposures, or single outcomes influenced by multiple behavioural and genetic exposures. Such studies also enable investigation of gene–environment interactions between genes and disease that vary by levels of environmental exposures (Kirkwood and Sterne 2003).

Biobanks vary in size, complexity of exposures and range of outcomes captured during follow-up. Moreover, they typically collect and store biological samples, including plasma, DNA, red cells, whole blood or urine samples. Combined analyses of lifestyle, clinical measurements, biochemical and genetic exposures together with fatal and non-fatal disease outcomes are likely to be particularly informative, not only for hypothesis testing but also for hypothesis generating. Table 1.6 includes

**Table 1.6** Selected characteristics of major biobank studies from around the world

Study names	Sample size	Location of study	Response rate (%)	Follow-up duration (years)	Disease outcome
European Prospective Study of Diet and Cancer (EPIC)	484,000	Europe	66	20	NF/F
Mexico City Prospective Study (MCPS)	150,000	Mexico	30	18	F only
China Kadoorie Biobank (CKB)	512,000	China	30	10	NF/F
UK Biobank Study (UKB)	500,000	UK	5–6	8	NF/F
Million Veterans Project (MVP)	825,000	USA	13	13	NF/F

NF non-fatal, F fatal disease outcomes

selected characteristics of some of the established large biobanks that have been conducted in Europe, UK, China, Mexico and the USA, respectively. These large studies highlight some of the modern developments of biobank studies during the twenty-first century, including the, unique combination of large size with breadth and depth of the data collected.

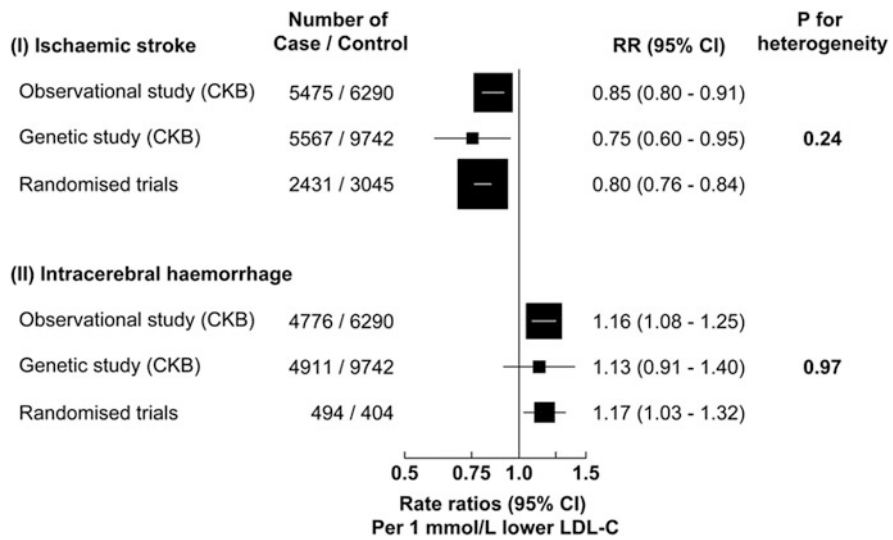
Reliable estimates of the relative risks of diseases in diverse populations are likely to be particularly informative. Biobank studies do not have to be representative of their source populations (i.e., despite having only a modest response rate), but if they are sufficiently large, their relative risks associated with major risk factors can be generalizable to a wide range of population subgroups (Batty et al. 2020; Chen et al. 2020). Heterogeneity between the different populations studied is important and combined analysis of large studies conducted in multiple diverse populations can assess a much wider range of exposures than is possible in any single population study.

There is a strategic need for prospective studies assessing effects of common exposures such as tobacco smoking, alcohol consumption, adiposity, blood lipids, blood pressure and diabetes in different populations as findings of such studies have a worldwide relevance. The CKB is a prospective study of 512,000 adults aged 30–79 years who were enrolled during 2004–2008 from 10 areas (5 rural and 5 urban) in China. After 10 years of follow-up, there were >50 K deaths and >one million hospitalizations for over 1300 different diseases including 60 K strokes, 30 K cancers and 50 K IHD. These exposures and disease outcome data are being complemented by large-scale genetics and other omics assays. Thus, prospective studies conducted in China can be immensely informative about important exposures (such as alcohol consumption, blood pressure or blood lipids) for diseases that are difficult to study in Western populations (such as different subtypes of stroke), but are still highly relevant to major NCDs worldwide (Sun et al. 2019; Millwood et al. 2019).

For some public health questions, the concordance of results of both observational studies and Mendelian randomization studies within the same population supported by results of meta-analyses of randomized trials in other populations can greatly strengthen causal inference and guide public health policies for prevention. For example, Fig. 1.13 shows the concordant results for equal and opposite associations of a 1 mmol/L lower LDL-cholesterol for ischaemic and intracerebral haemorrhage obtained from prospective studies, genetic studies and meta-analyses of randomized trials.

## 1.7 Accurate Reporting of Epidemiological Studies

As a large number of epidemiological studies are being published each year, the ability to assimilate, critically appraise, and apply such research findings from population studies is important. Epidemiologists typically follow reporting guidelines, such as Strengthening the Reporting of Observational studies in Epidemiology



**Fig. 1.13** Adjusted rate ratios (RR) for risk of ischaemic stroke and intracerebral haemorrhage associated with 1 mmol/L lower LDL-C in observational and genetic analyses in CKB, and in randomized trials of LDL-C-lowering drug treatment in Western populations (Sun et al. 2019)

(STROBE) criteria when reporting their findings, which include relevant details of design, inclusion criteria for participants, details of bias, confounding and specification of the statistical analyses used which have enhanced the quality and accuracy of reporting of epidemiological studies (von Elm et al. 2007).

## 1.8 Summary

This chapter provided an overview of the key concepts and principles that underpin the design and conduct of classical epidemiological and modern biobank studies. It also described the strengths and limitations of different types of epidemiological studies, with relevant examples from traditional and contemporary studies. Moreover, it highlighted the need for extremely large studies, importance of strict control of chance, bias and confounding, and outlined approaches for establishing causal inference for some key associations. Prospective cohort studies have undergone substantial transformation in recent decades, with significant increase in the size of population studied, the range of exposures collected, and the number and details of the disease outcomes studied. Advances in technology have enabled many biobank studies to undertake large-scale or cohort-wide multi-omics assays to supplement genome-wide association analyses that will greatly improve our understanding of the causes, prevention and treatment of major chronic diseases. However, the establishment and management of large biobank studies require careful planning and

attention to detail to overcome the substantial practical challenges, which will be discussed in subsequent chapters in this handbook.

## References

- Baigent C, Peto R, Gray R, Staplin N, Parish S, Collins R. Large-scale randomized evidence: trials and meta-analyses of trials. Oxford: Oxford University Press; 2020.
- Batty GD, Gale CR, Kivimiki M, Deary IJ, Bell S. Comparison of associations in UK Biobank against prospective general population-based studies with conventional response rates; prospective cohort studies and individual-level meta-analyses. *BMJ*. 2020;12(368):M131.
- Chen Z, Keech A, Collins R, Slavin B, Chen J, Campbell TC, Peto R. Prolonged hepatitis B virus and association between low blood cholesterol concentration and liver cancer. *BMJ*. 1993;306:890–4. <https://doi.org/10.1136/bmj.306.6882.890>.
- Chen Z, Emberson J, Collins R. Strategic need for large prospective studies in different populations. *JAMA*. 2020;323:309–10. <https://doi.org/10.1001/jama.2019.19736>.
- 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 perspective studies. *Am J Epidemiol*. 1999;150:341–53.
- Collins R, MacMahon S. Reliable assessment of the effects of treatment on mortality and major morbidity, I: clinical trials. *Lancet*. 2001;357:373–80.
- Floud S, Simpson RF, Balkwill A, Brown A, Goodill A, Gallacher J, Sudlow C, Harris P, Hofman A, Parish S, Reeves GK, Green J, Peto R, Beral V. Body mass index, diet, physical activity and the incidence of dementia in 1 million UK women. *Neurology*. 2020;94(2):e123–32.
- Hennekens CH, Buring JE. *Epidemiology in medicine*. Boston: Little, Brown and Co; 1987.
- Keys A. Seven countries; a multivariate analysis of death and coronary heart disease. Cambridge: Harvard University Press; 1980.
- Kyu HH, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, et al. Global, regional, and national disability-adjusted life-years (DALYs) for 359 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet*. 2018;392(10159):1859–922.
- Kirkwood BR, Sterne JAC. *Essential medical statistics*: Wiley-Blackwell; 2003.
- Lawlor DA, Smith GD, Bruckdorfer KR, Kundu D, Ebrahim S. Those confounded vitamins: what can we learn from the differences between observational versus randomised trial evidence? *Lancet*. 2004;363:1724–7.
- Liu B, Floud S, Pirre K, Green J, Peto R, Beral V for the Million Women Study collaborators. Does happiness itself directly affect mortality? The prospective UK Million Women study. *Lancet*. 2016;387:874–81.
- MacMahon S, Peto R, Cutler J, Collins R, Sorlie P, Neaton J, Abbott R, Godwin J, Dyer A, Stamler J. Blood pressure, stroke and coronary heart disease. Part I, prolonged differences in blood pressure: prospective observation studies corrected for regression dilution bias. *Lancet*. 1990;335:765–74.
- Millwood IY, Walters RG, Mei XW, et al. Conventional and genetic evidence on alcohol and vascular disease aetiology: a prospective study of 500 000 men and women in China. *Lancet*. 2019;393:1831–42.
- Ouellette S, Tassé AM. P(3)G - 10 years of toolbuilding: from the population biobank to the clinic. *Appl Transl Genom*. 2014;3:36–40.
- Prospective Studies Collaboration. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *Lancet*. 2002;360:1903–13.



- Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol.* 2003;32:1–22.
- Sun L, Clarke R, Bennett D, et al. Causal associations of blood lipids with risk of ischemic stroke and intracerebral hemorrhage in Chinese adults. *Nat Med.* 2019;25:569–74.
- Vandenbroucke JP, Broadbent A, Pearce N. Causality and causal inference in epidemiology: the need for a pluralistic approach. *Int J Epidemiol.* 2016;45:1776–86.
- von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet.* 2007;370:1453–7.

# Chapter 2

## Design, Implementation, and Management of Biobank Studies



Zhengming Chen

### Contents

2.1	Introduction .....	28
2.2	Plan of Investigation .....	29
2.3	Ethical and Legal Considerations .....	38
2.4	Study Protocol .....	39
2.5	IT Infrastructure and Systems .....	40
2.6	Quality Assurance Framework .....	41
2.7	Study Assessment Centre .....	43
2.8	Central Biobank Infrastructure .....	46
2.9	Study Organisation and Oversight .....	47
2.10	Summary .....	49
	References .....	49

**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](mailto:zhengming.chen@ndph.ox.ac.uk)

© Springer Nature Singapore Pte Ltd. 2020

Z. Chen (ed.), *Population Biobank Studies: A Practical Guide*,

[https://doi.org/10.1007/978-981-15-7666-9\\_2](https://doi.org/10.1007/978-981-15-7666-9_2)

27

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., anti-hypertensive 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	<ul style="list-style-type: none"> <li>• A variety of fractions: plasma, serum, white cell, red cells, peripheral blood lymphocytes</li> <li>• Wide range of biomolecules: DNA, RNA, proteins, small molecules</li> <li>• Wide coverage of physiological functions: genome, proteome, and metabolome, haematological parameters</li> <li>• Suitable for a wide range of assay technologies</li> <li>• Ease and low cost of collection</li> </ul>
Urine	<ul style="list-style-type: none"> <li>• Wide range of biomolecules: proteins, analytes (including pharmaceuticals)</li> <li>• Wide coverage of physiological functions: proteome and metabolome (including gut microbiome)</li> <li>• Suitable for many assay technologies</li> <li>• Ease and low cost of collection</li> </ul>

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  $\mu$ l 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

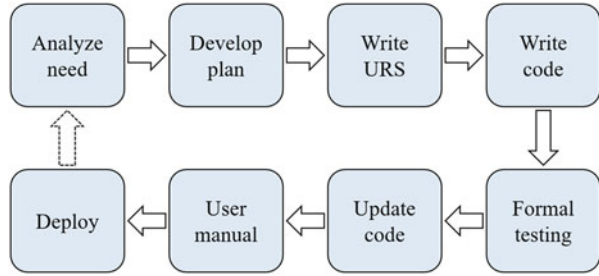
statement indicating even if the subject agrees to participate they may opt out for certain items (e.g., without providing biological samples) and can withdraw at any time without giving any specific reasons, although individuals who have expressed such wishes before joining the study should not be encouraged 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.

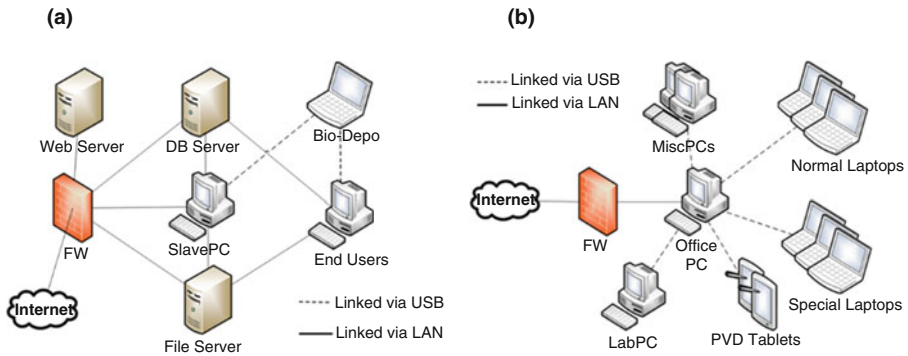
**Fig. 2.1** Standard process for developing IT software in biobank studies



## 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).

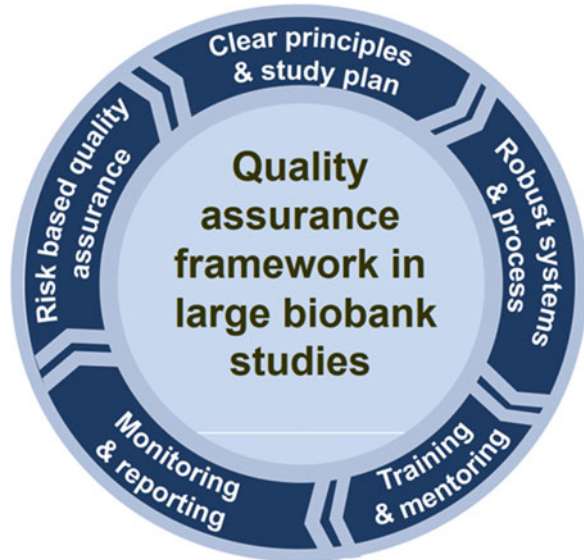
IT hardware devices that may be considered in biobank studies will vary, ranging from mobile phones, desktop computers, laptops, tablets, to servers, or even large cloud-based supercomputing and storage facilities. Very often different types of IT devices will be needed in the same study to meet different 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



### ***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.

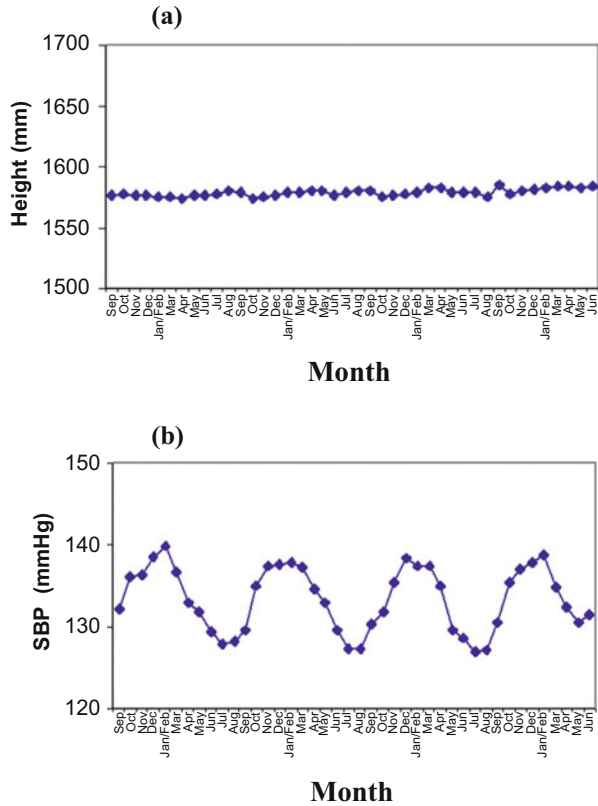
### **2.6.3 Study Monitoring**

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.

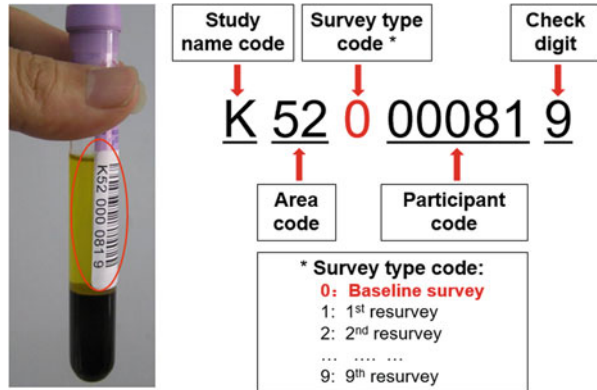
### ***2.7.2 Survey Team and Training***

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  $-130^{\circ}\text{C}$ , 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  $-80^{\circ}\text{C}$  freezers. Genomic DNA, especially when amplified, should be stored at  $-20^{\circ}\text{C}$ . 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.

# Chapter 3

## Planning, Organisation, and Management of Fieldwork in Biobank Studies



Ka Hung Chan, Kin Bong Hubert Lam, and Huaidong Du

### Contents

3.1 Introduction .....	52
3.2 Preparations for the Fieldwork .....	53
3.3 Initiating and Launching Fieldwork .....	65
3.4 Day-to-Day Management of Fieldwork .....	68
3.5 Monitoring and Quality Control .....	73
3.6 Summary .....	74
References .....	75

**Abstract** The importance of fieldwork in population health research has not diminished in the era of big data. Contemporary prospective biobank studies tend to collect large amounts of exposure and health outcome data from a large number of participants. Although disease outcome data are often obtained through linkages with registries and hospital records, the exposure data are generally collected directly from the participants, through questionnaire surveys (e.g., in-person, telephone, postal, or online), physical measurements, and collection and assays of biological samples (e.g., blood, urine, and saliva) at baseline recruitment and at other time points. There are many legal, logistic, and practical challenges in establishing large biobank studies, many of which are related to the fieldwork. Careful planning and organisation, efficient coordination and management, and effective implementation and monitoring are critical to ensure a successful launch and smooth operation of fieldwork. This chapter delineates key considerations and procedures involved in planning and organising fieldwork in biobank studies. Moreover, it describes certain novel methodological approaches relevant for the smooth operation of assessment centres and acquisition of high quality data. The general approaches described should also be applicable to other population-based studies (e.g., cross-sectional surveys and case–control studies).

---

K. H. Chan · K. B. H. Lam · H. Du (✉)

Big Data Institute Building, Nuffield Department of Population Health, Old Road Campus  
University of Oxford, Oxford, Oxfordshire, UK

e-mail: [huaidong.du@ndph.ox.ac.uk](mailto:huaidong.du@ndph.ox.ac.uk)

**Keywords** Prospective studies · Biobanks · Fieldwork · Data collection · Training · Quality assurance

## Abbreviations

API	Application programming interface
BP	Blood pressure
cIMT	Carotid artery intima media thickness
CKB	China Kadoorie Biobank
DNA	Deoxyribonucleic acid
ECG	Electrocardiogram
GE	General electric
ID	Identification number
IOP	Intra-ocular pressure
IT	Information technology
PWV	Pulse wave velocity
SOP	Standard operation procedure

## 3.1 Introduction

Compared with traditional study designs, contemporary prospective biobank studies tend to be much bigger in sample size and more comprehensive in terms of the exposures and health outcomes assessed. The large scale combined with the breadth and depth of data collected in biobank studies enables reliable assessment of the potential effects of many established and emerging risk factors for a wide range of diseases. While certain personal information can be collected indirectly via routine administrative and other records (e.g., occupational records), most exposure-related data need to be obtained directly from the participants themselves during initial enrolment at the baseline survey and subsequent periodic resurveys through interviews, physical measurements (including the use of wearable devices), and collection and assays of biological samples. In order to ensure that the data collected are of the highest possible quality, fieldwork must be carefully planned and implemented according to the study protocol and procedures, and include comprehensive management and quality assurance frameworks and systems. A detailed fieldwork plan should be developed to cover a wide range of practical and logistical issues, including assessment centres, data collection procedures, timelines, community engagement, survey team and training, pilot studies and subsequent scaling-up, integration of different components (e.g., information technology [IT] systems) and monitoring. This chapter outlines the key considerations and practical procedures necessary for organising fieldwork in large biobank studies, complementing the general introduction already covered in Chap. 2. Other aspects of fieldwork

relating to long-term follow-up for health outcomes and adjudication of fatal and non-fatal disease events will be discussed specifically in Chaps. 5 and 6.

## 3.2 Preparations for the Fieldwork

The success of fieldwork in prospective studies requires meticulous planning and careful preparation, typically involving many years of effort prior to commencing recruitment (see Chap. 2). Multiple social, economic, and environmental factors (Box 3.1) could influence the ways in which fieldwork is undertaken. The requirements and challenges may differ substantially in high- versus low-resource settings, and in urban versus rural areas within a country. For example, in high-resource settings, use of telephone, postal, or online questionnaires may be both preferable and more cost-effective than face-to-face interviews when recruiting participants from geographically diverse areas. However, in communities with limited access to telecommunication and/or poor literacy levels face-to-face interviews may be the only viable option. On the other hand, interviewer-administered questionnaire surveys may facilitate a higher response rate in many low-resource settings, even though they are labour-intensive and logistically challenging. Moreover, collection of biological samples and physical measurements can only be managed effectively through direct contact with the participants. In prospective studies, approaches used in the initial baseline survey and subsequent resurveys may also differ. For example, a high proportion of the UK Biobank participants have been periodically invited to complete a detailed dietary questionnaire online, which was not included at the initial baseline survey (Sudlow et al. 2015). Whilst undertaking the preparatory work, a number of fundamental aspects common to all study settings will need to be considered and are outlined and discussed below.

### **Box 3.1 Contextual Factors to Be Considered When Preparing the Fieldwork**

- Local laws and regulations.
- Cultural norms, religions, and values.
- Level of urbanisation and developmental status.
- Social network structure.
- Healthcare system and infrastructure.
- Physical environment (e.g., residential environment, climate).
- Distance to study assessment centre (e.g., road transportation, courier services).

### 3.2.1 Study Instruments

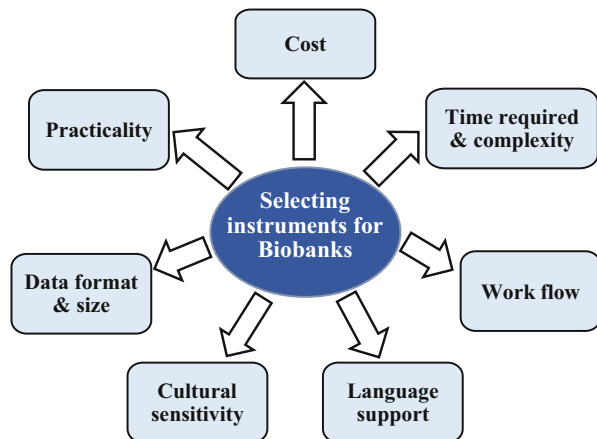
Study instruments refer to the tools for systematic data collection from study participants and include questionnaires and devices for physical measurements. The choice and design (where applicable) of instruments demand a balance between need and feasibility (see Chap. 2). Apart from perceived scientific values, several important practical issues should also be considered when deciding on the study instruments (Fig. 3.1).

#### (a) Questionnaires

Different biobank studies will have different objectives and different data collection plans, making study-specific questionnaires necessary. Typically, it is preferable to combine (and adapt) multiple existing questionnaires, ideally validated previously in the same study population. Such an approach would enable investigators to develop bespoke questionnaires that can reliably assess a wide range of demographic, socio-economic, and lifestyle factors relevant to local research need. Apart from conducting a thorough literature review, the design of questionnaires may also involve focus-group discussions with members of the target population and local stakeholders (researchers, healthcare professionals, and policymakers) to acquire in-depth understanding of the population characteristics and potential cultural sensitivity, as well as the acceptability and practicality of different topics or questions. Such interaction is crucial to inform the design of questionnaires.

Before finalising the questionnaire design, field-based pilot tests must be conducted in the target study population to understand the relevance of individual questions, the appropriateness of the questionnaire sequence, length, layout, precise wording (particularly for questionnaires translated from other languages), and response ranges (see Box 3.2, [Armstrong 2008] & Sect. 3.3.4).

**Fig. 3.1** Key practical considerations for selecting study instruments



### (b) Devices

For physical measurements, the most important scientific attributes to be considered are the validity and reliability (or reproducibility) of the data collected. Ideally, only devices with the highest validity and reliability are to be used, but they tend to be more expensive and require highly-trained technicians and sometimes special infrastructure to operate and therefore may not be feasible for large-scale studies, especially those with multiple sites running simultaneously. Practical issues, such as funding, population characteristics, physical infrastructure, ease of use, manpower, and utility (water, electricity) supply must also be considered. For example, ultrasonography of calcaneus may be preferable to the “gold standard” dual-energy X-ray absorptiometry for assessing bone health in large-scale biobank studies. This is because temporary assessment centres are often not located within healthcare facilities and the costs of providing X-ray based measurements which meet legal/regulatory requirements may be too prohibitive (Li et al. 2019).

Devices with operating interfaces in the local language are preferred to facilitate training and data collection. Where possible, it is more desirable to choose devices that can be controlled by computers directly through an application programming interface (API; generally provided by the device manufacturer) to bypass the need to export data saved on the device using proprietary software, thereby allowing a seamless data collection process. Other important factors to be considered include costs (procurement, running, and maintenance), applicability (in terms of the suitability of deployment in the field and the complexity to operate), acceptability (by both participants and fieldwork staff: invasiveness, comfort, convenience), and the expertise/training required. Provision for alternative data collection methods may be necessary for participants with particular conditions. For instance, an electronic weighing scale should be available for individuals with a pacemaker or pregnant women where contraindications for bio-impedance measurements exist. Likewise, mercury sphygmomanometers should be used in the small proportion of participants for whom blood pressure readings might not be successfully obtained using an automatic electronic sphygmomanometer (e.g., those extremely underweight).

#### **Box 3.2 Different Aspects of Validity of a Questionnaire**

- *Face validity*—the perceived accuracy of an instrument based on the professional judgement of relevant experts.
- *Construct validity*—the extent to which an instrument or specific question accurately measures the targeted metric, usually benchmarking certain gold standards.
- *Convergent validity*—the extent to which measures that theoretically should be related to each other are correlated; also considered as a type of construct validity.

(continued)

**Box 3.2** (continued)

- *Discriminant validity*—the extent to which measures that theoretically should be unrelated to each other are not correlated; also considered as a type of construct validity.
- *Known-groups validity*—the extent to which an instrument can differentiate between two (or more) groups known to differ regarding the variable of interest (e.g., smokers versus non-smokers).

### 3.2.2 *Collection of Biological Samples*

One of the defining features of biobank studies is the collection of biological samples, often stored long-term for future analysis. Biobanks typically collect a wide range of biological samples (most commonly blood, urine, saliva, and faeces), but some studies also collect environmental samples (such as air, water, or soil). The choice of samples to be collected is chiefly determined by the study objectives, the available budget, and the need for future-proofing, but other determinants including convenience and feasibility, acceptability by participants, ease of long-term storage also come into play (see Chap. 4).

**Box 3.3 Key Considerations of Blood Sample Collection During Fieldwork**

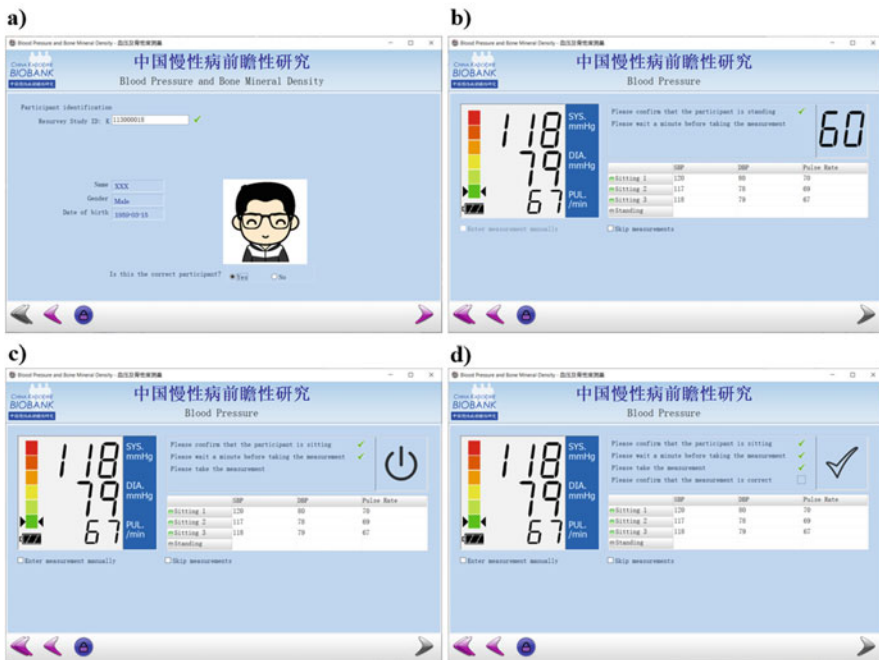
- Sample requirements (e.g., fasting, postprandial)—influence the flow of the fieldwork and at which stage the sample should be collected.
- Sample volume—dictates the storage requirements and the number and type of assays that can be done subsequently.
- Sample processing—some assays require specific fractions of the whole blood, which need to be separated within a short timeframe.
- The use of preservatives and/or anticoagulants—some of these may not be compatible with certain assays, whereas specific inhibitors must be used for some tests.

Most biobanks collect blood samples in order to conduct biomarker (e.g., biochemistry, hormones, proteomics, or metabolomics) and genetic analyses (e.g., genotyping or DNA sequencing) for associations with disease outcomes. However, the future use of the blood samples should be clearly planned at the preparation stage, because it will have a major influence on the fieldwork procedures (see Box 3.3). Central to any ethical approval, participants must be properly informed about the detailed plan for biological sample collection as well as future uses, including any feedback of assay results.

### 3.2.3 IT Support

Having adequate and appropriate IT support is one of the cornerstones of a successful biobank study. IT contributes to every aspect of the fieldwork, but most commonly it is incorporated in the data collection processes. Electronic data collection removes the need for labour-intensive and time-consuming paper record processing and minimises manual data entry errors and missing data through built-in and automated real-time checks for logic, error, and range. Moreover, through automated encrypted data transfer to secured servers in data warehouses, it also minimises the risk of data breaches and loss.

Figure 3.2 provides screen captures of the bespoke software system for blood pressure measurement in the China Kadoorie Biobank (CKB) study (Chen et al. 2011). It first confirms the participant’s identity by checking the synchronised photo image captured at the initial registration against that of the participant attending the station and then measures and transfers blood pressure readings from the device to the study computer via Bluetooth connection. The system also specifies the time interval between repeat measurements, ensuring adherence to Standard Operating Procedure (SOP).



**Fig. 3.2** Example of bespoke software to record blood pressure measurements in CKB. (a) Verifying participant identity, (b) Ensuring rest before measurement, (c) Taking measurement after proper rest, (d) Confirming & saving results into computer



Compared to paper-based questionnaires, electronic questionnaires can be highly flexible, with questions automatically popping out or bypassed based on the individual's response to previous questions (see Chap. 2 examples in Fig. 3.3). For example, questions on the amount of alcohol consumed would not be shown for those who report never drinking. Integrated logic and range checking allows real-time quality control of data.

IT development is a relatively expensive and time-consuming process. However, in large studies the benefits offered by high efficiency of electronic data collection, together with improved data quality and flexibility in data collection and quality assurance greatly outweigh the time and monetary costs of software development, hardware procurement, and maintenance. Where possible, the research team should include expertise in IT and software development to support the project long-term because: (1) bespoke systems can then be designed to meet the specific requirements of the study; (2) technical problems can be quickly resolved; and (3) long-term sustainability of the biobank will require ongoing maintenance of the IT infrastructure. Outsourcing may be a cost-effective alternative where in-house IT expertise is unavailable or very limited. In all circumstances, IT development requires close collaboration between investigators, software developers and, where appropriate, manufacturers of the devices (see Chap. 7).

### ***3.2.4 Development of Standard Operating Procedures***

Once the decision on the data collection methods is made, the development of SOPs should begin. SOPs aim to provide clear instructions for fieldworkers to perform their tasks in a uniform way, thus minimising systematic and random errors due to variability between staff and over time. SOPs are study-specific and should be written by researchers who are familiar with the subject matter and fieldwork, and where appropriate, in collaboration with the IT programmers to facilitate the development of relevant software/systems. Additional comments and requests may emerge during the development of SOPs, requiring modification and refinement of the procedures (and software). Therefore, proper version control of documents is important for record tracking and future reference. If the local language is not supported by the devices and/or off-the-shelf software packages are used, the SOPs must include translations of the text in the user interface and all dialogue boxes.

A good SOP should have all the key qualities listed in Box 3.4. It should provide clear and succinct step-by-step instructions on the procedures and all necessary information for troubleshooting and reporting errors. Instructions must be simple, preferably in the form of bullet points, supplemented by graphical illustrations such

1. Background information

What is your current occupation?

**Red box: pop-up questions only shown after the option "Retired" is selected in the 1<sup>st</sup> question**

- Agricultural and related worker
- Factory worker
- Administrator/ manager
- Professional/ technical
- Sales and service worker
- Retired
- House wife/ husband
- Self-employed
- Unemployed
- Other or not stated

What was your last occupation before you retired?

- Agricultural and related worker
- Factory worker
- Administrator/ manager
- Professional/ technical
- Sales and service worker
- House wife/ husband
- Self-employed
- Unemployed
- Other or not stated

Why did you retire?

- Reaching retirement age
- Health related (excluding injuries)
- Other reason

How many people live together as a family in the household?

persons

How often do you interact socially with people outside your household (e.g. by talking to people in person or on the telephone or other media)?

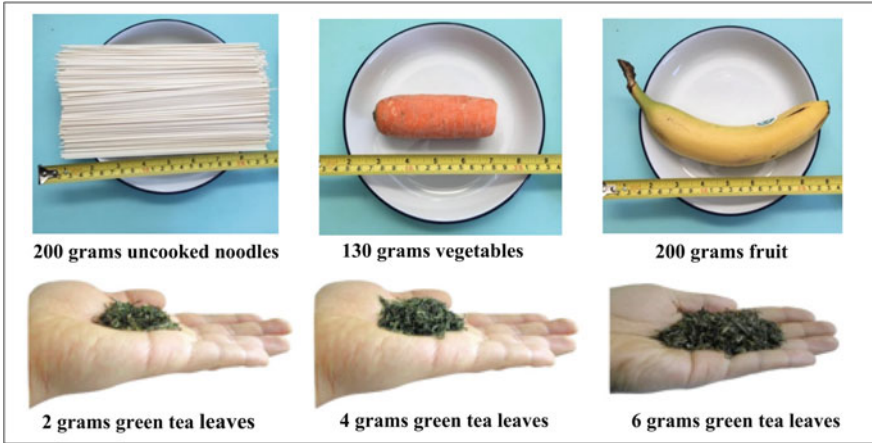
- Daily or almost every day
- A few times a week
- A few times a month
- Never or rarely

What is your current marital status?

- Married
- Widowed
- Separated/ divorced
- Never married

< Prev Help Next >

Fig. 3.3 Example of the laptop-based electronic questionnaire in CKB



**Fig. 3.4** Illustrative images to assist questionnaire interview on food consumption

as flowcharts and screen captures. User manuals from the manufacturer for the devices used for physical measurements or collection of biological samples may need to be translated or adapted (typically simplified) to suit study-specific needs. Where there are contraindications or exceptions for certain procedures, the alternative approach must be clearly described in the SOPs so that fieldwork staff know what they should do in such situation. For example, for participants who are unable to stand up straight, their arm span can be a proxy for standing height.

#### **Box 3.4 Key Qualities of Standard Operating Procedures (SOPs)**

- Clarity.
- Accuracy.
- Conciseness.
- Step-by-step instructions.
- Tailored to local context.
- Abundant graphics to facilitate comprehension and memorisation.
- Written in simple language easily understandable by fieldworkers.

SOPs for interviews should include clear guidelines on how to ask the questions (e.g., intonation), definitions and elaboration of individual questions, and suggested responses for common queries by the participants. Printed leaflets with relevant photographic guides may be a helpful aid when the questions are difficult/complex (e.g., type of medications taken) or involve quantification (e.g., amount of foods or alcohol consumed) (Fig. 3.4).

Often the SOPs for different procedures are collated as a single fieldwork manual, to serve as a universal reference document for every single procedure required for the fieldwork. A proper index is essential even though the SOPs are likely to be listed in

a logical order resembling that envisaged for the fieldwork. In the CKB, a fieldwork manual was produced for the baseline survey and each subsequent resurvey, which included over two dozens of SOPs detailing the procedures of setting up assessment centres, conducting interviews, undertaking a variety of physical measurements, collection and management of biological samples, and storage and transfer of data (Box 3.5).

### ***3.2.5 Study Assessment Centres***

A study assessment centre is the place where participants attend to undertake assessments such as interview, physical measurements, and sample collection. The number of assessment centres to be established is contingent on the anticipated sample size, locations, and planned timeline for recruitment and the budget available. To encourage participation, the centre should be located conveniently within the study area with good transport links. It should not have disturbing features such as background noise and unpleasant smells. Sometimes, assessment centres are located in healthcare facilities or other public premises (e.g., government buildings, community centres, schools), but this depends largely on local circumstances and may not be feasible especially if the expected duration of recruitment is long. Researchers may have to set up purpose-built centres in temporary or semi-permanent structures (e.g., gazebo, shipping container).

Irrespective of the type of facility that houses the study assessment centre, all essential amenities, including electricity, toilet facilities, and heating (if appropriate) must be available. Ideally, the centre should be located on the ground floor or in a multi-storey building with elevator access to facilitate attendance of elderly or disabled participants and movement of study equipment and consumables. The space should be large enough to be partitioned into different stations to suit different needs (e.g., private room for certain tests such as ECG). Where possible, all stations should be on the same floor, allowing the participants to navigate the process easily and facilitating communication (including possible wireless data transfer) between stations. In settings with a limited choice of venues, researchers may need to modify the design of the fieldwork or choice of instruments to accommodate the physical constraints. For example, for venues with insufficient/inconvenient toilet facilities, on-site urine sample collection may not be feasible and the participants may be instructed to collect the sample at home instead.

**Box 3.5 List of SOPs for Fieldwork of the Second Resurvey in CKB**

- Establishing regional offices, laboratories, and local storage facilities for fieldwork-related equipment and consumables;
- Establishing assessment centres;
- Establishing, training, and managing field survey teams;
- Management of study equipment and consumables;
- Informed consent and participant registration;
- Management of consent forms (from collection to collation and archiving for long-term storage);
- Anthropometric measurement (including standing and sitting height, waist and hip circumferences);
- Measurement of body composition (body weight and fat percentage) using a Tanita bio-impedance device;
- Assessment of blood pressure and heart rate;
- Measurement of handgrip strength using a Jamar hydraulic hand dynamometer;
- Measurement of exhaled carbon monoxide using a MicroCO meter;
- Measurement of lung function using a Vitalograph spirometer;
- Carotid artery intima media thickness (cIMT) and plaque examination using a Panasonic CardioHealth Station;
- Assessment of pulse wave velocity (PWV) using a Pulse Trace PCA2 device;
- Assessment of bone mineral density using a GE Achilles EXP II bone ultrasonometer;
- 12-lead electrocardiogram (ECG) examination using Mortara ELI 250c instrument;
- Guidance for conducting face-to-face questionnaire interview;
- Blood sample collection and on-site measurement for blood glucose and lipids;
- Urine sample collection and benchtop measurement of urinary biomarkers;
- Temperature and relative humidity monitoring;
- Data monitoring (completeness and quality);
- Unexpected events logging;
- Data transfer (i.e., from assessment centre to local office, and then to national/international collaborating centre);
- On-site bio-sample management (including temporary storage at assessment centre and shipment from assessment centre to local laboratory);
- Bio-sample aliquoting, transfer (i.e., from local laboratory to central sample storage facility of the CKB study), long-term storage of and logging;
- Non-respondent survey.

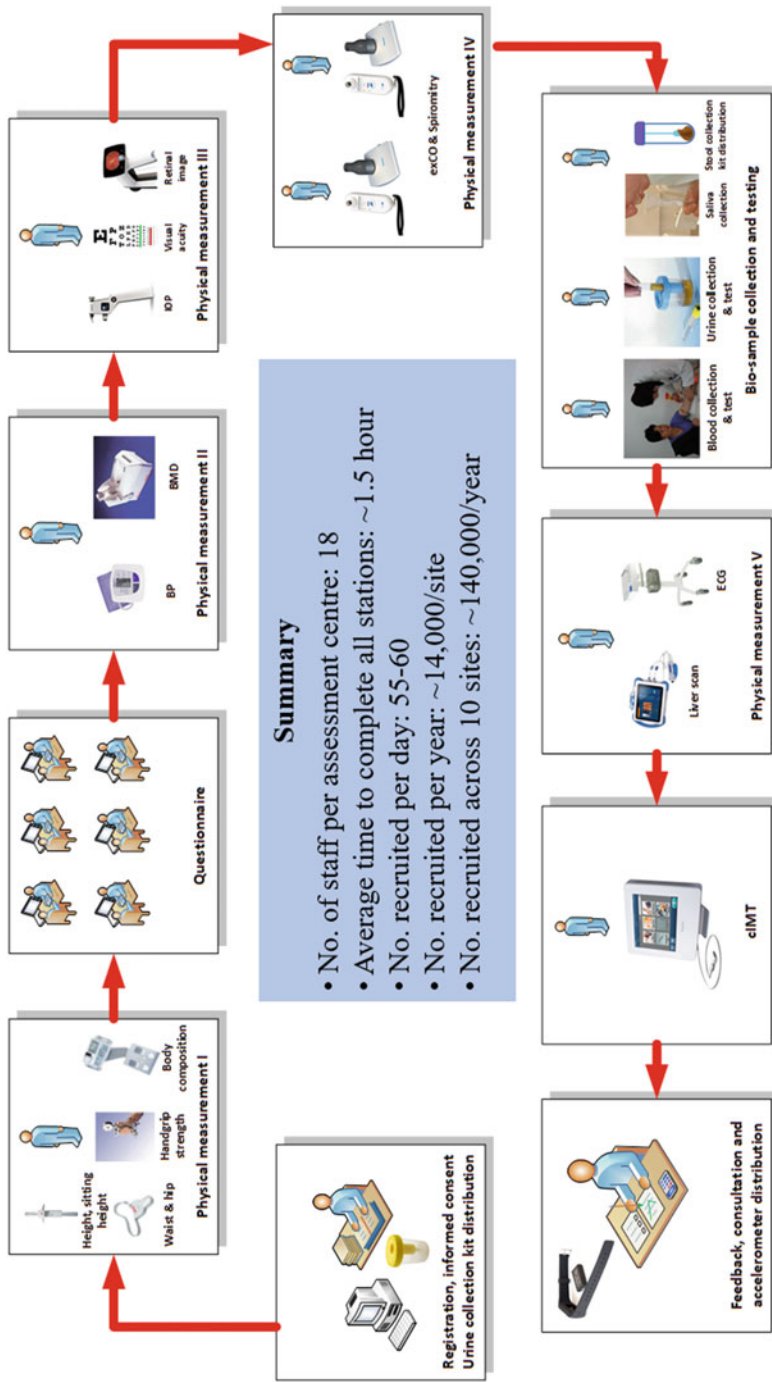


Fig. 3.5 Workflow, station names, and number of staff for each station in CKB fieldwork

### ***3.2.6 Optimising Fieldwork Flow Across Different Stations***

Once the scope of the fieldwork is determined, the sequence of different procedures should be designed to standardise data collection and maximise data quality and efficiency. In the workflow example shown in Fig. 3.5, several issues have been taken into consideration. First, the measurements of blood pressure (BP) and intra-ocular pressure (IOP) should take place before spirometry to avoid fluctuations and artefacts introduced by forced expiration. Second, since height and weight readings are required for spirometry and liver scans, anthropometric measurements should be undertaken first. Third, urine specimen cups are distributed at the beginning to maximise the success rate for collection of urine samples during the visit and minimise waiting time. Finally, liver scan and ECG measurements can be taken at a single station as both require the removal of clothing above the waist and measurement in a lying position.

The workflow should be carefully planned to benefit from better IT integration. If the computers across different stations are networked, data collected from one station can be synchronised with the other stations to avoid unnecessary data entry and ensure the correct order of measurements is followed. For example, the computer at the spirometry station could verify whether the anthropometric, blood pressure, and eye measurements have already been undertaken. Sex, age, and height could also be automatically imported into the spirometry software. An alternative to networked computers is achieved by providing each participant with a USB key upon registration that can both record and transfer data across different stations. A carefully planned workflow also enables a more realistic projection of the time needed for the entire visit, as well as maximising the total number of participants that could be surveyed per day with the given number of staff and physical infrastructure available.

### ***3.2.7 Establishing the Fieldwork Team***

It is always preferable to recruit the fieldwork team staff well in advance according to the study plan, timeline, and job specifications (see Chap. 2). While it is possible for researchers who design and lead the project to undertake the fieldwork themselves, it is often preferable for large biobank studies involving multiple diverse study sites to recruit the fieldwork team, including fieldwork manager/coordinator, locally. The possession of local knowledge (environment, culture, language) is essential for the effective and efficient implementation of the fieldwork. Fieldwork staff should have the appropriate educational qualifications and the relevant clinical/technical experience required for certain specific roles (e.g., phlebotomy, spirometry, electrocardiography, and ultrasound imaging). However, it may not always be possible to recruit such candidates and specific in-house training is sometimes required. Staff from local hospitals or healthcare providers, or university/college students in the region

could make excellent fieldworkers, but their availability throughout the study duration cannot be guaranteed.

### **3.3 Initiating and Launching Fieldwork**

Earlier in this chapter, overarching considerations in the planning of fieldwork have been discussed. As the launch of the fieldwork becomes closer, researchers will need to shift their focus to more localised aspects, including social mobilisation, staff training, piloting, setting-up of assessment centres, team management and capacity building, instrument and consumable management, and sample and data management.

#### ***3.3.1 Community Engagement and Social Mobilisation***

Engagement with local communities from planning to implementation is key to an effective recruitment campaign. In many low-resource settings, communities are very cohesive and establishing trust from the key local leaders (politicians, religious leaders, or elders in the community) is a prerequisite for access to individuals in such communities. Regardless of the setting, researchers should first approach community representatives (e.g., resident associations) to understand the local needs and concerns, and to explain the purpose and importance of the study, including how it might benefit the local community (e.g., understanding the local disease burden and healthcare needs). From these discussions, the approaches and contents of publicity activities, in addition to formal invitations and enrolment processes, can be tailored to the local communities. Key considerations include socio-demographic-economic characteristics and social network structures of the community, available financial and human resources, and national and international ethical guidelines.

Based on the knowledge acquired and rapport built over the community engagement process, a sustained and widespread course of social mobilisation should be organised to promote the study and encourage participation throughout the fieldwork. The local community leaders (heads of villages and chairs of community committees) may be recruited to assist the study team by distributing leaflets and posters as well as informally communicating with local residents (e.g., through telephone calls, face-to-face conversations, and community-wide broadcast/social media). During the fieldwork, the leaders may continue to act as mediators between the fieldwork team and the community. Individual participants are also encouraged to spread the news and invite their friends and relatives who are potentially eligible for participation in the study, creating a bottom-up momentum of participation over a relatively short period of time. Since most large biobank studies involve multiple communities (e.g., CKB covered 100–150 communities in each of the ten survey sites), the order and timing of community engagement should be carefully planned



such that high levels of engagement can be achieved and maintained in the communities under survey or to be surveyed very soon. This in turn requires accurate estimation of the time needed to complete the survey in each community, which would vary by the population size, response rate, manpower, and season.

### ***3.3.2 Setting Up Assessment Centres***

The location of study assessment centres should be decided at the planning stage (see Chap. 2 and Sect. 3.2.5). For most biobank studies, especially those conducted in resource-limited settings, temporary assessment centres tend to be co-located within existing facilities. Under such circumstances, one should consider the content and requirements of the fieldwork and whether the layout of the available space can be adapted for the fieldwork purpose (e.g., private space required for procedures such as ECG and liver scans may be created by temporary partitions); the existing infrastructure (e.g., provision of utility, accessibility); and the estimated timeline of the fieldwork and the availability of the space (both size and duration). Site visits are very important for investigators to understand what is available.

A carefully designed and set-up assessment centre is essential to the smooth implementation of fieldwork. The configuration of different data collection stations should be easily followed by study participants according to a pre-specified workflow (see Sect. 3.2.6). For example, the registration station should be located near the entrance of the centre to capture all attending participants upon arrival. Typically, there should be a spacious waiting area with an adequate supply of chairs, drinking water/refreshments/newspapers. To further engage the study participants, banners, posters, and video clips about the study and the fieldwork can be displayed throughout the assessment centre. The size of the allocated space for each station should be determined by the number of staff members, facilities needed (e.g., bed for ECG measurement), and the consumables and devices involved. For stations where bottlenecks are more likely to emerge (e.g., due to the complexity and time needed for the procedures), there should be flexibility to allow extra staff to be deployed. Practicality and safety are also important factors to be considered. For example, eye measurements have to be done in a dimmed environment, dedicated space is needed for collecting and processing biological samples to avoid accidents (such as needle-prick injuries) and contamination, and stations requiring electricity should be located near the mains sockets.

### ***3.3.3 On-Site Training***

The purpose of on-site training is to allow survey staff members to familiarise themselves with their tasks and to ensure they perform such tasks exactly as specified in the SOPs. Even though each fieldwork staff member may be assigned to one to

two specific roles, it would be optimal to ensure that all key tasks can be performed by multiple staff members to enable fieldwork to be carried out without interruption when particular individuals are unavailable for various reasons (e.g., sick leave, holiday).

The SOPs or fieldwork manuals should be made available to staff members well in advance before training begins to allow them to get familiarised with the study design and training requirements. The duration of on-site training will vary depending on the study objectives, range, and complexity of the data collection. Typically, the training should include formal presentations, highlighting the study objectives, key design principles, practical procedures involved in the fieldwork, and issues covered in SOPs. These should then be followed by practical training sessions, which enable trainees to practise study procedures and also provide an understanding of the perspectives of study participants through role play exercises. Aptitude tests could be used to identify study staff who are particularly suitable for specific tasks. Throughout the training sessions, various formal and informal tests should be undertaken to identify gaps and enhance training wherever necessary.

Use of multimedia materials such as video clips can facilitate training for more complex procedures such as carotid ultrasound scans and the setting up of air pollution sensors. With widespread availability of smart phones (even in low-resource settings), video clips can be readily circulated to study staff based in different locations (Arku et al. 2018). Where necessary, certain video clips can also be shown to participants at assessment centres to facilitate certain measurement processes (e.g., spirometry) that may require good cooperation of participants.

The fieldwork training providers should keep detailed records about the background experience of each fieldworker and identify their strengths and weaknesses relevant to their roles over the course of training. Using this information, fieldwork coordinator and team leader can assign fieldworkers who are better at communication and engagement to the interview stations, and those with clinical experience to procedures such as phlebotomy.

### **3.3.4 *Piloting***

Piloting is a crucial step prior to launching the main study, but its importance is often under-appreciated. For a large biobank study, multiple rounds of piloting may be required at different stages of development to test a wide range of practical procedures and systems, not just those related to data collection. Pilot studies may involve a small number of volunteers from local communities to go through all the stations and follow procedures as close to the actual ones as possible, in order to help identify any unexpected problems and streamline the organisation of fieldwork. These pilot studies should be incorporated as the final part of the on-site staff training mentioned above. Trainers and supervisors should observe each individual fieldworker closely throughout the process and provide specific and constructive

feedback about their performance on a daily basis. Any new issues identified should be recorded and addressed in future training sessions.

### ***3.3.5 Supervised Launch of Fieldwork***

In order to maximise the utility of the training process and sustain the momentum among the fieldwork team, the actual fieldwork should be launched immediately after the end of training and piloting with the help of senior research and fieldwork staff, who normally provide the training. This may be considered as a “supervised launch” phase, when researchers and IT programmers are present in the field to closely supervise and support the team. Such a “supervised launch” phase may last for several days, with the recruitment rate set a relatively low level, say 30–50% of target daily rate, to allow the fieldwork team to become familiar with their tasks gradually in a real-world setting. It would also allow researchers and IT programmers to identify new issues not envisaged previously. To start with, the assessment centre may operate only in the morning, with the afternoon devoted to debriefing and discussion sessions, and if necessary further training. It is also the time for fine tuning the execution of certain procedures or the setup of the assessment centre, and ironing out any logistical and IT glitches. As the performance of the fieldwork team improves and becomes more independent, the pace of the fieldwork can be ramped up to the targeted rate.

## **3.4 Day-to-Day Management of Fieldwork**

Biobank studies tend to take place in multiple sites and last for a prolonged period of time. Effective day-to-day management of fieldwork is important and involves materials supply, equipment calibration and maintenance, staff training and mentoring, monitoring of workload and recruitment rate, and handling and processing of data and samples. Dedicated members of the study team should be assigned and trained to conduct specific tasks. Depending on the local conditions and particulars of fieldwork, different approaches can be undertaken but the following aspects should be considered.

### ***3.4.1 Survey Team Management***

One key aspect of fieldwork coordination lies in human resources management. It is important to devise team building activities to motivate staff members and foster cooperation and mutual trust. Formal or informal assessment can also be undertaken to evaluate and monitor the quality of work undertaken by team members, and

certain rewarding systems (e.g., outstanding performance award) can be useful to enhance the commitment and quality of work. One way to facilitate data quality control is to assign each member a unique staff ID number used in the purpose-built electronic data collection system, so that data quality and consistency can be monitored. This would be particularly important for new staff and for those who temporarily provide cover for absence of certain fieldworkers due to holiday and/or sick leave.

### ***3.4.2 Device Calibration and Maintenance***

In large biobank studies, a wide range of different devices is often utilised to collect data. To ensure sustained reliability of these devices, calibration and regular maintenance are critical. Certain devices have to be calibrated each day (or time) before use, while others may only require maintenance on a monthly or annual basis as recommended by the manufacturer. For the former, it is especially important that the calibration routine be clearly specified in the SOPs. While having built-in calibration mandates will make it impossible to omit the calibration task (as it cannot be bypassed), there may still be a need to plan where and how to save the calibration data which might be needed for subsequent data analysis (for example, as a correction factor). For devices where only periodic calibration is needed (as recommended by the manufacturer), a system must be in place to notify the due date (ideally integrated in the electronic fieldwork management system), if it is not automatically prompted in the user interface or software. As a minimum requirement, the following information should be appended to the SOPs: (1) scheduled calibration frequencies; (2) consumables needed; (3) steps of calibration; and (4) alternative strategies in case of failure of calibration. All technicians operating the devices must be trained to carry out the calibration procedures and to document and report any such activities and problems discovered clearly and accurately. Logs of calibration must be kept for future reference and should be checked periodically by fieldwork coordinators for compliance.

Similarly, the maintenance schedule and procedures of different devices should also be documented systematically, preferably in an electronic inventory management system. Any information relevant to the device should be entered to this system, including (1) authorisation of the purchase; (2) purchase order and invoice; (3) the year when the device was purchased; (4) shipment records (including customs clearance, where relevant); (5) manuals and instructions for calibration and maintenance; (6) consumables needed; (7) number purchased and their warranty period; (8) contact details of the manufacturer (or distributor); (9) the lead time for replacement; (10) current locations and condition (fully functional, requiring repair); and (11) the scheduled dates of service/calibration and maintenance.

### ***3.4.3 Supply of Consumables***

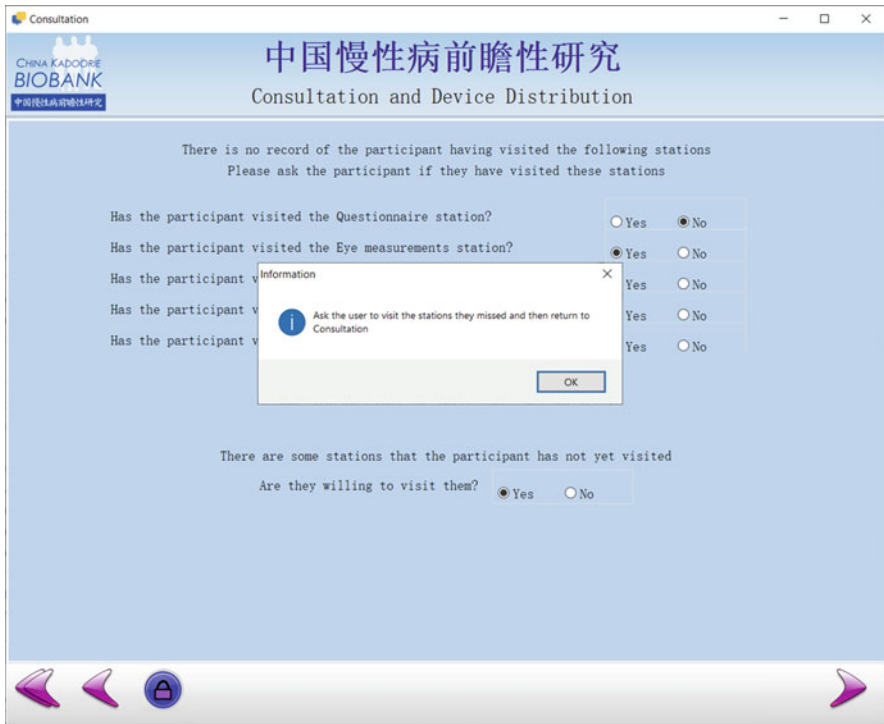
In planning large biobank studies, much of the effort is likely to be invested in the data collection instruments and devices, from decision to procurement to shipment. It is important that the consumables (e.g., syringes, sample collection tubes, mouth-piece of spirometers, on-site assay reagents) be given the same attention. As most studies will span over several years, a stable supply of consumables has to be achieved at all cost to minimise disruption to the study. It is essential to accurately estimate the amount of consumables needed each cycle (day or week), so that sufficient but not excessive amount of consumables can be prepared and shipped to the assessment centre regularly. In some countries where extreme weather conditions are frequent, the assessment centres may need to stock up with more consumables in case delivery is interrupted. The fieldwork coordinator need to have an inventory list and check the stock regularly. Monitoring the rate of consumption of consumables (given the number of participants recruited) is also essential, as it captures potentially unnecessary and excessive wastage of materials.

The same electronic system used to manage devices can also be used to manage and monitor consumables. For instance, the system can record purchase dates, sources, and quantities, where they have been distributed and used, down to the assessment centre or fieldwork team member level. The system can be linked with recruitment data to enable automated notification messages being generated to prompt restocking of consumables when the stock level is below certain pre-set thresholds.

### ***3.4.4 Participant Invitation and Management***

Against the backdrop of social mobilisation, the participant invitation process must be carefully managed. The key considerations include (1) the expected response rate; (2) the number of participants that can be assessed each day; (3) organisational strategies to ensure a seamless workflow; and (4) the communication plan and the mechanism to handle complaints or special requirements during the fieldwork. With appropriate social mobilisation, it is possible to see a large turnout of potential participants as well as non-eligible individuals, especially in cohesive communities in rural areas. It is prudent for the fieldwork coordinators to work with community leaders in keeping participation at a sufficiently high but manageable level. In the case where residential records are available, invitation letters can be sent to potential participants at different times so that the assessment centre can be run at the optimal capacity; if such contact information is not available and invitations must be sent through informal channels, staggering by geographical areas can be done where possible.

Since long-term follow-up is essential to all prospective biobank studies, it is important to recruit participants who are likely to be stable and long-term residents of



**Fig. 3.6** Automated check for completeness of fieldwork data collection in CKB

the community to minimise loss-to-follow-up. To better understand the response rate as well as the potential differences between participants and the non-respondents, researchers should, where possible, aim to acquire some information on the socio-demographic characteristics of non-respondents, and reasons for non-participation, and compare them with that of the target population.

To safeguard the privacy of participants and to protect confidentiality of personal information, anonymised unique identification numbers (study IDs) must be used throughout the study. During fieldwork, participants may be asked to wear a wristband labelled with study ID because these cannot be easily swapped between participants. At each assessment station, the study ID number should be carefully checked (and entered using a barcode reader). If the stations are inter-connected through networked computers, as is the case in CKB, photographs of participants may be taken upon registration and then synchronised across the internal computer network in the assessment centre. The photograph can be displayed with the study ID at any station, as an additional confirmation of the identity and ensure correct linkages with the data captured at particular stations. Before the participants leave the assessment centre, there should be a final check on the completeness of data collection. As shown in Fig. 3.6, a computer programme can check automatically if all stations have been completed when the participant's study ID is scanned at the

final station. If any station has been missed, a pop-up window will prompt the fieldworker to send the participant to the appropriate station. This is also a good time to distribute additional data or sample collection instruments (e.g., wearable accelerometers, food diaries, stool sample collection kits) to eligible participants.

### ***3.4.5 Biological Sample Management***

Detailed strategies for sample collection and management (e.g., processing, shipment, and long-term storage) are discussed in Chap. 4. The tubes or containers used to collect biological samples must be labelled clearly with barcodes of participants' unique study IDs, to enable effective matching of sample-related data with other personal information. It is often necessary for the samples (e.g., blood and urine) to be analysed centrally in an accredited laboratory, rather than on-site. Therefore, to prevent/minimise degradation, once collected at the assessment centre, the samples must be stored in a low temperature environment (on ice or 4 °C refrigerator) before being transported to the laboratory chilled, and analysed/aliquoted as soon as possible (ideally within 12 h from collection). Multiple aliquots of biological samples in cryovial tubes should be transferred in different batches and stored in different freezers and cryogenic tanks to prevent accidental loss. To manage and keep track of the large quantity of samples in a biobank study reliably over a long period of time, it is essential to use a computerised sample logging and tracking system (see Chap. 4).

### ***3.4.6 Data Management***

Once collected, the data should be carefully stored and managed to maintain security and confidentiality. Tailored systems should be established to manage data transfer and processing both locally and centrally. All electronic data collected from the field should be stored and transferred in password-protected USB keys and encrypted in computers and hard-drives. Data collected on paper (e.g., signed consent forms) should be shipped to local offices by a dedicated member of staff for scanning and storage as confidential data. All data are then transferred to multiple firewall-protected central databases for integration via a secured IT infrastructure (not email or unencrypted cloud service). Personally identifiable data must be stored separately from other health-related data such that the analytical databases can be fully anonymised. Data stored in the fieldwork assessment centre (computers and devices) should be deleted once the data have been transferred to the local/central offices. At the end of the fieldwork, all study computers should be securely erased to remove any trace of the participant data (see Chap. 8).

## 3.5 Monitoring and Quality Control

Quality assurance is particularly important to ensure the quality of all data collected and research outputs over subsequent decades. Apart from SOPs and IT systems, staff training and fieldwork monitoring are the key elements for quality assurance programmes (see Chap. 2).

### 3.5.1 Study Monitoring

Fieldwork monitoring can be considered as an extension of staff training and mentoring, whereby the performance of individual team members is closely monitored throughout the course of the study. Supervisors should aim to provide regular feedback, mitigate the problems identified, prevent future occurrence, and thus improve fieldwork quality. Monitoring can be broadly categorised into active or passive approaches, but a mixture of both approaches is often necessary.

Active monitoring may include direct auditing of the performance of fieldworkers, through audio (or video, if feasible and acceptable) recordings made through study computers. This type of quality control procedure may require ethical approval and should be explicitly included in the study consent forms and discussed during formal staff training. Other types of active monitoring may include careful review of a small random samples of specific data (e.g., appropriateness of conversation with participants from recording) and images (e.g., carotid artery ultrasound images) collected daily by experienced staff or trained technicians (e.g., sonographers) in the coordinating centre and collaborating hospitals.

Passive monitoring typically involves the automated mechanisms to analyse and assess the data collected by relevant staff members (e.g., quality of lung function manoeuvres), which could help identify further training needs and areas for improvement. More sophisticated checks may involve statistical monitoring for reproducibility of repeated measurements or for comparison with some expected (e.g., centre-specific average) or target values. Descriptive statistical analyses can be performed on the data collected daily by each staff member to check for discrepancies, outliers, and irregularities. For instance, they may cover (1) frequency of male current smokers; (2) distributions of measured blood pressure levels; (3) the last digit numbers for automated measurements (e.g., blood pressure) that may identify possible malfunction of instruments or poor data collection; (4) mean time to complete questionnaire interview by different staff; and (5) mean time delay from sample collection to processing. Compared to active monitoring, passive monitoring has a major advantage of being more cost-efficient, because such approaches can “react” immediately upon data entry or sample collection to correct any error without the need for human intervention. see Chap. 2



### **3.5.2 Contingency Planning**

Contingency plans should be considered to minimise the effects of unexpected events or interruptions to any component of the fieldwork. In many low- and middle-income countries, sudden power outage is especially common, so it is essential to have an emergency power supply (e.g., petrol/diesel generator) in the assessment centre. To avoid accidental loss, data must be regularly backed-up, daily or more frequently, to multiple secured locations. Redundancy should be implemented in all aspects of the fieldwork. At least one readily deployable backup spare for each physical device or computer must be available. Moreover, there must be a contingency plan in place to allow for the malfunctioning of key equipment. For example, if a wireless network is used to facilitate data synchronisation across different stations, paper-based forms should be available as a backup so that data collected from one station (e.g., body weight and height assessed at body composition station) could be available for the other stations (e.g., lung function assessment) in case of network failure.

Similarly, additional staff should be trained and be available on stand-by in the event of illness or unplanned leave in the fieldwork team. Where it is impossible to recruit additional personnel, members of the team should be trained for multiple roles with periodic refresher training and practice. At least one member of the study team in the assessment centre should be a qualified first aider. Where necessary, medically-qualified staff may be employed to monitor drug administration (e.g., bronchodilator in spirometry) and to perform phlebotomy. They may also provide resuscitation and treatment in the event of a medical emergency. In settings where extreme weather conditions are common (or at least not rare), researchers may also need to consider the risks of natural disasters, such as typhoons, flooding, or wildfires, during which the fieldwork should be halted and relevant disaster and weather emergency protocols should be launched to protect or retrieve study staff, participants, equipment, and data. Certain indemnity insurance policies should be arranged to protect staff and study assets in case of accidents.

## **3.6 Summary**

This chapter provides a general overview of the key considerations required in planning and undertaking field work for large prospective biobank studies. Most of the issues covered and practical procedures described reflect in part the process involved in setting up the large CKB which recruited over 0.5 million participants from 10 geographically diverse localities in less than 4 years. While the importance of careful planning and effective coordination should never be underestimated, development of novel approaches and methodologies, especially robust and bespoke IT systems, to support and manage the field work from data collection, staff and material management, to quality assurance and monitoring, play a critical role in

ensuring smooth operation of the field work. Different strategies and approaches may be developed by other population health studies with different study designs, sample size, data collection, settings, and populations. Irrespective of the strategies adopted, careful planning, effective management, and comprehensive monitoring, aided by robust IT systems, are vital to ensure success of all such studies.

## References

- Arku RE, Birch A, Shupler M, Yusuf S, Hystad P, Brauer M. Characterizing exposure to household air pollution within the Prospective Urban Rural Epidemiology (PURE) study. *Environ Int.* 2018;114:307–17.
- Armstrong BK. Validity and reliability studies. In: White E, Armstrong BK, Saracci R, editors. *Principles of exposure measurement in epidemiology: collecting, evaluating, and improving measures of disease risk factors*. 2nd ed. Oxford: Oxford University Press; 2008.
- 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(6):1652–66.
- Li X, Qiao Y, Yu C, Guo Y, Bian Z, Yang L, Chen Y, Yan S, Xie X, Huang D, Chen K, Chen Z, Lv J, Li L. Tea consumption and bone health in Chinese adults: a population-based study. *Osteoporosis Int.* 2019;30:333–41.
- Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* 2015;12:e1001779.

# Chapter 4

## Collection, Processing, and Management of Biological Samples in Biobank Studies



Iona Y. Millwood and Robin G. Walters

### Contents

4.1 Introduction .....	78
4.2 Study Design and Planning .....	79
4.3 Sample Collection .....	82
4.4 Sample Processing .....	85
4.5 Sample Shipment .....	90
4.6 Sample Storage and Retrieval .....	92
4.7 Strategies for Sample Analysis .....	93
4.8 Monitoring and Troubleshooting .....	95
4.9 Summary .....	96
References .....	97

**Abstract** Prospective biobanks provide opportunities for investigating the contributions of a wide range of lifestyle, environmental, and genetic factors to risk, aetiology, and prediction of many different diseases. A growing array of high throughput technologies, capable of measuring of hundreds, thousands, or millions of biochemical and genetic factors, can support investigation of the relationship of such factors to disease risk or risk factors, inform on aspects of behaviour or lifestyle that are otherwise difficult to measure reliably (e.g., diet), and enable the generation and testing of hypotheses concerning the causes of disease. Thus, integral to any biobank study is the collection, processing, and storage of biological samples for use in such technologies. This chapter describes the main steps involved in these processes. There will be an emphasis on study design and developing procedures to ensure that the types of samples collected will be suitable for the intended analyses, and that they are processed, transported, and stored under conditions that will preserve their integrity and allow them to be used for a range of future research

---

I. Y. Millwood (✉) · R. G. Walters  
Big Data Institute Building, Nuffield Department of Population Health, Old Road Campus  
University of Oxford, Oxford, UK  
e-mail: [iona.millwood@ndph.ox.ac.uk](mailto:iona.millwood@ndph.ox.ac.uk)

purposes. Requirements for sample linkage and security will also be considered. These considerations are not limited to prospective studies, but are relevant to any study (e.g., retrospectively recruited case-control cohorts) that involves the collection and storage of biological samples from large numbers of individuals.

**Keywords** Biobanks · Cohort studies · Biological sample · Assay · Storage · Processing · Linkage

## Abbreviations

CKB	China Kadoorie Biobank
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
EDTA	Ethylenediaminetetraacetic acid
PST	Plasma separator tube
RNA	Ribonucleic acid
SOP	Standard Operating Procedure
SST	Serum separator tube

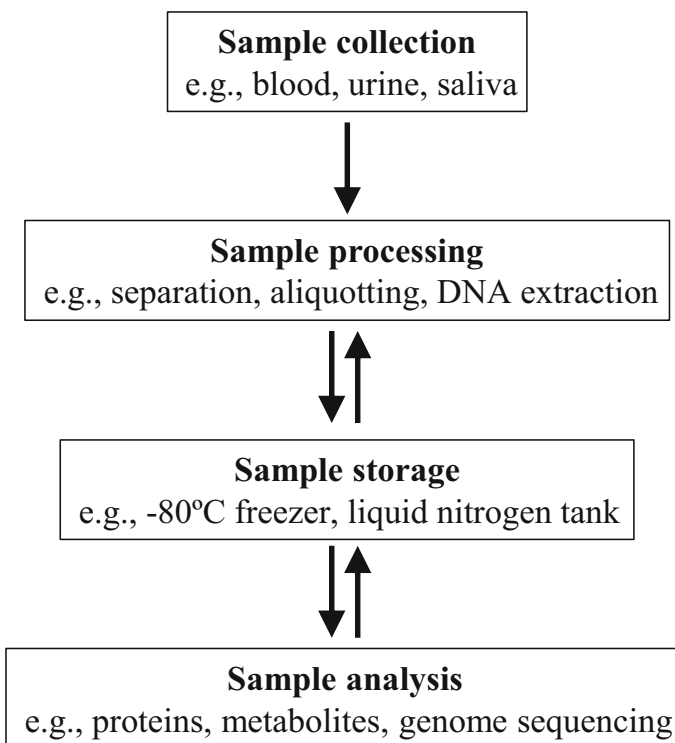
## 4.1 Introduction

Prospective biobanks provide opportunities for investigating the contributions of a wide range of lifestyle, environmental, and genetic factors to risk, aetiology, and prediction of many different diseases. To enable exploitation of current and future high throughput technologies, an integral part of a biobank study is the collection, processing, and storage of appropriate biological samples, to enable the measurement of many different cellular, genetic, biochemical, and other biomarkers, thereby greatly expanding the range of clinical and epidemiological research questions that can be investigated using the biobank (Chen et al. 2011; Sudlow et al. 2015). In particular, measuring and assessing biomarkers can help researchers investigate and understand disease aetiology and pathological mechanisms, improve clinical tools for risk prediction, diagnosis and prognosis, and identify potential therapeutic targets. Large-scale sample collection in biobanks, generally conducted among the general population, is a huge endeavour and must be carefully planned, and be convenient, cost-effective, and acceptable to participants. Samples must be collected, processed, and stored under standardised and controlled conditions which maintain their integrity. To achieve that, detailed protocols should be developed and piloted to ensure procedures are reliable and practical, and that samples are accurately and robustly linked to the individuals from whom they were collected. Samples can be used for the analysis of biomarkers at the time of collection or shortly afterwards, or may be stored long-term for future research. The types of sample collected, and the conditions and tubes in which they are collected,

processed, and stored, will determine which biomarkers can be analysed over the course of the study. Strategies for biological sample collection should aim to maximise the amount of high-quality data that will be obtained from the limited biological resources, and to allow for future, as yet unforeseen, analysis of stored samples. This will ensure the biobank can achieve its immediate research goals, can participate in collaborative efforts with other biobanks, and is able to adapt to future advances in technology.

## 4.2 Study Design and Planning

At the time the biobank study is being designed it is essential to give careful consideration to both the overall biological sample strategy and the detailed procedures for each stage of sample collection, processing, storage, retrieval, and analysis (see Fig. 4.1). Literature review, together with consultation between the study investigators, stakeholders, and the wider scientific community, should be undertaken to plan which samples should be collected, whether the proposal will be



**Fig. 4.1** Overview of biological sample handling for biobanks

acceptable to participants, whether it is practical given the available resources and infrastructure, and what are the main priorities for future sample analyses.

#### ***4.2.1 Approval for Collection***

The proposed biological sample strategy will need to be consistent with national and local regulations regarding sample collection and use, and to be approved by the relevant ethics committee(s), who will consider its scientific justification and any inconvenience and discomfort to participants. Study ethics applications should include a plan for the handling of incidental findings from sample analyses (e.g., genetic predisposition to disease). Participant information materials and consent procedures will need to explain the collection, storage, and potential future uses of the samples to be collected, so that participants are able to make an informed decision about their participation (see Chap. 2). Participants must be able to withdraw consent, including for the storage and use of their biological samples and any associated data.

#### ***4.2.2 Developing SOPs***

Once the overall study strategy is in place, detailed procedures (protocols or SOPs) should be developed and validated before beginning sample collection. These should cover every stage and describe every step of sample collection, processing, storage, retrieval, and analysis. It is important to conduct pilot studies of sample collection and processing, to ensure that procedures are clear, practical, easy to follow, can be consistently applied by many different staff across multiple study sites, produce samples that are fit for purpose, and that IT systems function correctly.

Processing of samples after collection, into separate aliquots or sub-fractions, should be streamlined to maintain sample integrity and minimise degradation and loss. This may be best conducted at a central laboratory soon after collection, but this has the disadvantage of requiring frequent (at least daily) transfer of samples from each recruitment site for central processing, with resulting delays before samples can be frozen for long-term storage. The relative benefits of centralised versus local processing will depend on the availability of appropriate facilities, a regular and reliable supply of ice for sample preservation, transport logistics, and the nature of the samples to be collected.

### 4.2.3 *Establishing Storage Infrastructure*

Key hardware and infrastructure should be identified and sourced well in advance. In particular, tubes for collection and storage of samples should be selected that not only are suitable for the required storage temperatures but can also be thawed and refrozen—possibly multiple times—without cracking or damage to their identifying labels. There should be sufficient fridge and/or freezer space in each location where samples will be stored, including local storage for samples prior to shipping to central sample repositories. Storage temperatures should be appropriate for long-term sample preservation, and a plan for transporting samples between different sites should be in place.

### 4.2.4 *Developing IT Systems*

Particular attention should be paid to developing IT systems and procedures to record all details of sample linkage, handling, and processing (see Chap. 7). In CKB, numerous IT applications are directly related to the collection and handling of samples (see Table 4.1), and additional systems record data relating to the contents of particular cryovials (e.g., sample volume remaining). Integrating electronic recording of information into sample collection, including use of barcode scanners to track and link participants and samples, can help to ensure that procedures are correctly followed and to reduce errors and omissions in data collection. Direct

**Table 4.1** Bespoke IT systems related to biological samples in CKB

Procedure	Program name	Details
Collection	BarcodeChecker	Check barcode printing on pre-labelled cryovials
	SampleLogging	Record the receipt of blood/urine packs from a day's clinic
	BrowsePacks	View the receipts of blood/urine packs from a day's clinic
Processing	SampleAli	Manage aliquoting of samples and record sample linkage
	Limousine	Synchronisation of study and laboratory databases
Movement	SampleTracking	Manage movement of sample boxes between centres
	Cryovials Handler	Transfer cryovials between sample boxes
	ExternalShipments	Manage sample shipments to external laboratories
	BoxHistory	Movement history of sample boxes between centres
Storage	StorageAreaConfig	Configuration of the sample storage capacity in a centre
	NccStore	Manage sample storage in freezers and nitrogen tanks
	IccStore & Ncc Sample Locator	Search and view graphically the physical storage of samples

communication between recruitment sites and a central coordinating centre can aid real-time monitoring of sample handling, while participant recruitment is ongoing, for early detection and rectification of problems.

### **4.3 Sample Collection**

In planning sample collection, it is important to consider the likely priorities of the biobank in terms of the diseases and risk factors to be investigated, which will inform the choice of biosamples to be collected and how they are processed and stored. In large population-based biobanks, it is essential that sample collection is practical, cost-effective, and able to be conducted at scale.

#### ***4.3.1 Settings for Sample Collection***

Sample collection also needs to be acceptable to participants. Particularly invasive procedures or collection of excessive sample amounts can discourage participation and so should be avoided unless clear benefits, in terms of additional information generated, can be demonstrated. Similarly, it should be considered whether fasting samples are needed, in light of inconvenience to participants, practical issues for clinic timing, and the value of the additional information gained (e.g., fasting could be important for a study with a focus on diabetes). If samples are non-fasting, data should be collected on how long it has been since the participants last ate or drank, since this will allow appropriate adjustment in analyses of the many biomarkers that are directly or indirectly influenced by food consumption.

Routine sample collection in community settings by study staff, for instance, during baseline recruitment, is generally successful and can achieve yields approaching 100% (e.g., CKB collected blood from 99.98% of participants). However, for some samples or measurements (e.g., stool, blood cell count) that are challenging to collect or analysis of which is complex, it may be decided to accept lower collection rates, for instance, by requesting samples from only a subset of participants. Samples collected at other times, for instance, if a second visit is required or if the participant collects the sample at home using kits supplied to them by the study, will inevitably have lower collection rates. As with other aspects of participant selection and recruitment, wherever samples are collected from a subset of participants, consideration should be given to how potential biases (e.g., due to particular subgroups not providing samples) will be avoided or mitigated.



**Table 4.2** Types of biological sample and potential uses

Sample type	Possible analyses
Blood fractions	
Whole blood	Haematology cell count, RNA (transcripts)
Red cells	HbA1c, fatty acids
Buffy coat	DNA (sequencing, genotyping, methylation)
Plasma or serum	Metabolites, proteins, antibodies, microorganisms
Urine	Metabolites, proteins, electrolytes, trace elements
Saliva	DNA (sequencing, genotyping, methylation), oral microbiome
Stool	Gut microbiome, occult blood
Other tissue (e.g., hair, nails)	Trace elements
Clinical specimens (e.g., tumour biopsy)	Disease specific factors, e.g., histopathology, somatic mutations

### 4.3.2 Sample Types

A prospective biobank study will typically collect whole blood and its fractions, which together enable measurement of a wide range of important genetic and non-genetic biomarkers. In addition to blood, it is common for biobanks to collect other sample types, such as urine, saliva, stool, or nail or hair clippings (see Table 4.2). Specialist studies (e.g., cancer biobanks) may collect tumour biopsies or other clinical specimens (Patil et al. 2018). Blood collection is relatively straightforward, and can provide a wealth of data. For example, a 10–20 mL blood sample will enable a wide range of analyses to be done, including measurement of circulating metabolites and proteins, and extraction of genomic DNA for genotyping or sequencing. As well as collecting samples for long-term storage, it may be decided to make immediate assays of some key biomarkers for which rapid testing is available (e.g., blood glucose, lipids), which may require collection of an additional sample.

### 4.3.3 Collection Procedures

Sample collection procedures should be clear and easy to follow, should minimise discomfort to participants, and should be standardised across multiple study sites, with consistency in procedures, equipment, and consumables. Equipment used should be regularly maintained and calibrated according to a clear schedule, with detailed record-keeping. Collection tubes or containers must be able to withstand routine handling, sample processing, and long-term storage conditions. All samples collected should be clearly labelled and securely and correctly linked to the participants from whom they derive throughout the duration of the study. The most efficient way to achieve this is through barcode labelling of sample tubes and recording sample linkage using IT systems (see Chap. 7), but this nevertheless

**Table 4.3** Samples collected at the UK Biobank baseline visit<sup>a</sup>

Sample	Vacutainer tube	Fraction
Blood	EDTA (9 mL × 2)	Plasma, buffy coat, red cells
Blood	Lithium heparin (PST)	Plasma
Blood	Silica clot activator (SST)	Serum
Blood	Acid citrate dextrose	DMSO blood
Blood	EDTA (4 mL)	Haematology (immediate)
Blood	Tempus tube	Whole blood (RNA)
Saliva	(collection vessel)	Mixed saliva sample
Urine	(collection vessel)	Urine

<sup>a</sup>Adapted, with permission, from Peakman and Elliott (2010)

requires careful planning to ensure that the recorded information is correct. In CKB, at study enrolment participants were assigned a unique study ID (8-digit code and associated barcode) and were given a sheet of printed labels with this study ID (but no other identifying information) to take to each workstation in the local study assessment centre. One of these labels was used for the blood collection tube, linking the blood sample to the participant.

Sample collection methods should be appropriate for the planned use of the samples, taking account of the need to stabilise samples to prevent degradation of certain biomarkers, while also considering the impact that preservatives may have on future assays. For instance, preservation of samples for future measurement of RNA transcripts requires use of special reagents that will degrade other biomarkers, and collection into EDTA tubes facilitates separation of buffy coat for DNA extraction and inhibits DNases but may interfere with some biomarker assays. It is also important to facilitate downstream sample processing and analysis, and pilot studies may be helpful in this. For example, the UK Biobank sample protocol underwent extensive piloting and validation before it was finalised (Elliott et al. 2008; Sudlow et al. 2015). At the baseline visit, blood (and saliva and urine) was collected into specialist vacutainer tubes with different preservatives and stabilisers, with different coloured lids (see Table 4.3), with multiple blood samples taken: for separation of plasma, anticoagulants were used (EDTA, lithium heparin), while for serum, clotting activators were used.

Samples may be collected by trained study staff, by health-care professionals or, with suitable instruction, by the participants themselves. Generally, samples should be collected at the clinic site at study enrolment or when other assessments are being conducted. Blood collection must be performed by staff trained in phlebotomy, whereas some sample types (e.g., urine, or saliva from mouth swabs) could be collected by the participant at the clinic or in their own home. In the case of biopsies and other clinical specimens, specialist clinicians will be responsible for sample collection in hospital settings.

### **4.3.4 *Minimising Potential Risks***

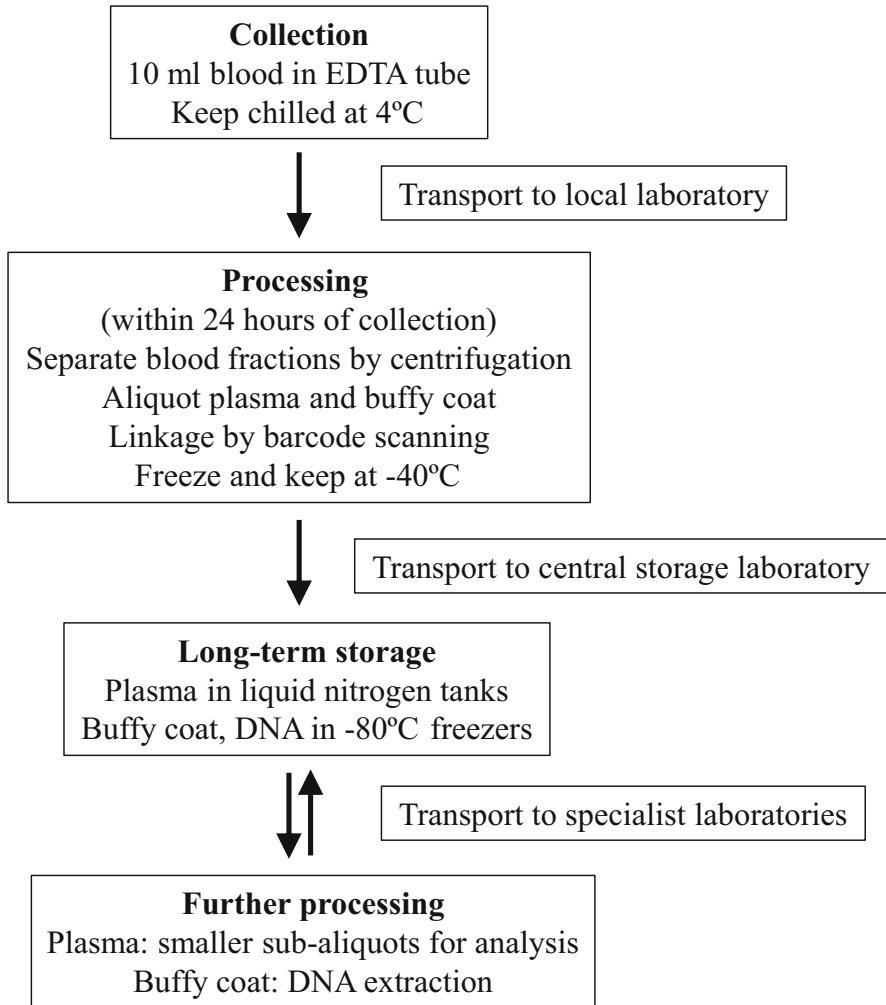
Biological sample collection may pose risks to participants (e.g., fainting during drawing of blood) and study staff (e.g., liquid nitrogen splashes). Appropriate health and safety procedures should be in place, with protective clothing such as gloves and lab coats worn by staff, as needed, and facilities for disposal of contaminated biological material and consumables. If it is planned to rapidly freeze samples in liquid nitrogen, in addition to appropriate equipment to protect participants and staff from burns it must be ensured that there is adequate ventilation. Needle stick injury during phlebotomy poses a significant risk, as blood samples may be infected with viruses such as HIV or HBV, and suitable protocols should be in place to minimise these risks and for treatment in case of such an event.

## **4.4 Sample Processing**

Biological samples will frequently need to undergo some form of processing before they are stored or analysed, and samples that are not immediately frozen at the time of collection will need to be kept chilled, either in a refrigerator at 4 °C and/or properly insulated box with ice and chilled packs (e.g., during transportation). Depending on the type of sample, different processes may be done at the time of collection at the clinic site, or centrally prior to storage, or at a later time when samples are analysed. It is important that sample integrity is preserved during processing, with key parameters such as duration and temperature, and number of freeze-thaw cycles, properly recorded and controlled to avoid degradation of samples and their analytes.

A key process is the creation of multiple aliquots from each sample, where the volume available and sample type permits it. This is important for sample preservation, and will provide more options for different analyses, avoid samples undergoing multiple freeze-thaw cycles, and allow the same sample to be stored in different locations (guarding against catastrophic failure of sample storage in one location). If at all possible, therefore, such aliquoting should be done immediately after sample collection. The number of aliquots and associated sample volumes which are made from each sample will depend on storage capacity and cost, as well as planned future use.

National and local occupational safety regulations should be followed when processing biological samples, which is often done using a biological safety cabinet, or following biohazard containment procedures appropriate to the level of risk (e.g., contamination from blood born viruses when handling whole blood). Unused biological material (e.g., that remaining after aliquots have been created) and contaminated containers and consumables must be disposed of safely according to regulations.



**Fig. 4.2** Overview of blood collection, processing and storage in CKB

#### 4.4.1 Blood Sample Processing

Most biobanks will collect one or more blood samples from participants. Blood should be chilled immediately after collection (or as otherwise required, e.g., serum samples should first be held at room temperature for 30 min, to allow clotting) and held in temporary refrigerated storage until processing. Processing at recruitment sites may be possible if facilities permit, but these may not have appropriate biohazard containment or sufficient freezer storage. Transfer to and processing at a

central facility will help to ensure consistency in sample handling. Blood samples will usually be separated into fractions (e.g., plasma, serum, buffy coat, red cells) by centrifugation, according to the desired downstream analyses and the type of vacutainer used for sample collection, each fraction then being transferred to one or more separate tubes before freezing for long-term storage.

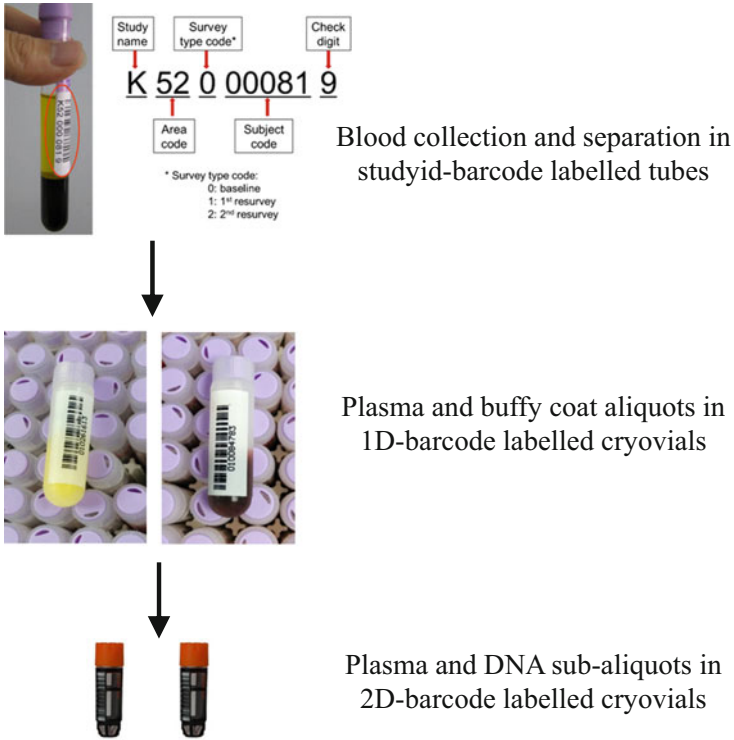
For CKB, which collected a single 10 mL EDTA blood sample from each participant, blood samples were chilled immediately after collection, transported by courier from the clinic site to a central laboratory at each of the ten study regions, and processed within 24 h of collection, with the mean time delay from collection to processing being ~10 h (see Fig. 4.2). Processing involved separating the fractions by centrifugation, followed by creation of one buffy coat and three plasma aliquots per sample, by manual pipetting. Aliquots were frozen and transported on dry ice to a central storage facility, for long-term storage at either  $-80^{\circ}\text{C}$  (buffy coat) or in liquid nitrogen (plasma). The plasma aliquots were stored at two different locations, providing security against sample loss.

Careful recording of sample linkage was a critical step at all stages of processing. Protocols at each stage were designed to minimise the chance of errors, and all cryovials used for plasma and buffy coat aliquots were pre-labelled with 1-dimensional (1-D) barcodes, with linkage recorded by barcode scanning of both the original blood sample tube and the new cryovials (see Fig. 4.3). Storage boxes for the cryovials were also labelled and scanned, and the position of each cryovial within each box was recorded, providing a second independent record of linkage that could protect against, e.g., damage to cryovial barcodes (see Fig. 4.4). To further ensure accurate linkage, box and cryovial barcodes included a check digit to protect against mis-scanning.

#### ***4.4.2 Sample Reformatting and Sub-Aliquoting***

Many biomarker assays, including several high throughput technologies, require only small volumes of plasma or serum, so that large numbers of different measurements are possible. However, repeated sampling of blood fraction aliquots for different assays, leading to multiple freeze-thaw cycles and opening and closing of cryovials, risks potential degradation and contamination of samples, along with accidental damages and loss. It can therefore be valuable to further divide the blood fractions produced at time of recruitment into multiple separate (potentially single-use) sub-aliquots, long-term storage of which will ensure that the study can exploit future, as-yet-unforeseen technologies. For such subsequent processing of stored samples, it is important to ensure that the methods used will reliably scale up to the large numbers of samples in a biobank without loss of yield/quality.

Sub-aliquoting can usefully be carried out at the time a sample is first used, thereby avoiding an unnecessary freeze-thaw cycle. Alternatively, it may be done systematically as a separate procedure, to prepare samples for future use. CKB adopted both these approaches—25,000 samples (1 mL) were divided into



**Fig. 4.3** Blood sample processing and barcodes in CKB

sub-aliquots (50–400  $\mu\text{L}$ ) at the time they were used for measurement of clinical biochemistry biomarkers, metabolomics, or proteomics, while a further 100,000 will be sub-aliquoted in preparation for future high throughput assays. This used 2-D barcoded cryovials in boxes of 96, which could be simultaneously scanned recording all cryovial positions in the box without the need for removing them from the box, allowing straightforward and robust tracking of sample movements and linkages.

#### 4.4.3 DNA Extraction

Other processing of blood samples will likely include extraction of genomic DNA from stored buffy coat, for which multiple different procedures are available. In CKB, the final decision on the method of extraction, an automated magnetic bead-based protocol, was only made after a series of pilot studies and the development of procedures and IT systems to track samples and monitor the quantity and quality of extracted DNA. On average, 1 mL buffy coat yielded about 90  $\mu\text{g}$  DNA, which was suspended in 400  $\mu\text{L}$  buffer and separated into two aliquots, one of which was

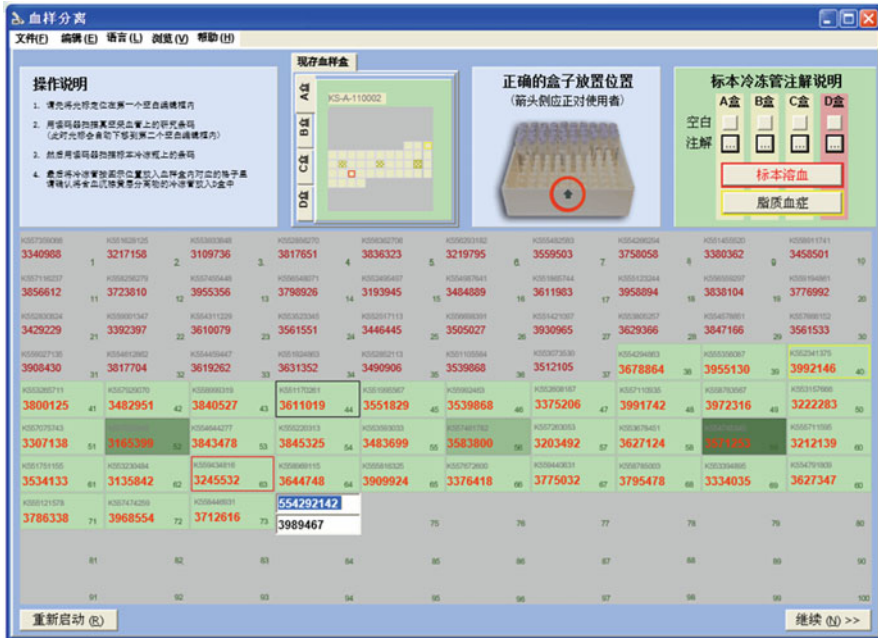


Fig. 4.4 Software linking study ID to cryovial barcode and position in storage boxes

placed in long-term storage. A third DNA aliquot was then made by diluting to a standardised concentration (50 ng/μL), to be used as a working stock for genotyping, sequencing, and other analyses.

Detailed procedures and IT systems were designed to ensure that linkage between the source tube and the resulting processed samples was robustly recorded. Buffy coat cryovials were scanned and their location in a 24-tube rack recorded prior to extraction and automated liquid handling, and then linked to the 2-D barcoded cryovials into which its DNA aliquots were transferred, which were themselves stored in boxes of 96, which were also labelled and scanned. As well as individual sample linkage, box number and sample position within the box were recorded for each cryovial.

### 4.4.4 Automation and Technology

Automation improves the quality and consistency of sample processing. For example, aliquoting on a large scale can be done efficiently and accurately using liquid handling robots, which can be programmed to aliquot set volumes into tubes in specific formats, recording data to accurately link samples, and log all steps of the process. Automated processing is faster than manual sample handling, more

consistent, and less prone to linkage errors. These systems can be very costly and it is not always going to be feasible to set up a new robotic liquid handling system, but if existing facilities are available, then studies should consider using them. However, errors that occur in automated sample handling will likely affect many samples. For example, leaking seals in the automated pipettes can result in samples being diluted with water, affecting results of measured samples, which may only be discovered when the data are used for analysis (UK Biobank 2019). It is therefore essential that liquid handling and other automated systems are carefully monitored, and regularly checked and serviced.

## 4.5 Sample Shipment

Samples will generally need to be transported, from local clinics or field sites, to central sites or laboratories for processing and storage. They may also be transported subsequently to different laboratories, sometimes in other countries, for further processing and specialist analyses.

### 4.5.1 *Mode of Transport and Packing*

The relative locations of the different study sites will determine the optimal mode of transport, and considerations should be made for delivery time and conditions, security, and cost. Different transport options include air, sea, rail, road, private vehicles or public transport, or by post. Some types of sample, or materials used to keep samples chilled or frozen (e.g., dry ice), may not be permitted on certain forms of transport, or may require specialist packaging. Shipments will usually be done by professional courier companies, or, if short distances are involved, delivery by study staff may be feasible if local laws permit and appropriate insurance is in place—for CKB, some transport from recruitment site to processing centre was performed by study staff. Samples collected by participants in their own home, or by local medical centres, may be sent by post to the processing or storage sites, as long as samples and packaging comply with local post regulations.

Samples must be packed securely in suitable storage boxes, cool bags, or padded envelopes, to protect them during shipment. In most cases, sample transport should be temperature controlled and monitored, to keep the samples chilled or frozen. The quantity of cooling materials, such as dry ice or freezer blocks, should be sufficient for the size of the shipment and the time of travel. Thermometers can be included in the packing cases for continuous monitoring of conditions.



### 4.5.2 Tracking and Risk Mitigation

All movement of biological samples should be closely monitored, so that the locations of samples can be identified at any particular time. In CKB, a “SampleTracking” software program was developed which recorded which samples and boxes were included in a shipment, collection and arrival times, and the condition of samples upon arrival (see Fig. 4.5). All samples were checked out of one site and checked into their destination, by barcode scanning of individual cryovials and/or boxes.

Loss, damage, or degradation of samples during shipment is a potential risk. For example, if a shipment is delayed *en route*, dry ice in the packaging may dissipate and samples may thaw. An important method of mitigating this risk is to ship different aliquots from the same sample in different shipment batches, to avoid loss of all the sample from a participant. Shipment of samples should be avoided during major public holidays, or when key staff are not available to closely monitor progress, and should be sent only at times when staff will be available (and are aware they are expected) to receive and appropriately handle them at their destination. Biological sample shipment must be done in accordance with relevant local or national regulations. The shipping agreement with courier companies will usually require documentation of any potential biological hazard posed by the samples, and these companies should have suitable insurance to cover their liability in the event of sample loss or damage, or in case of accident leading to exposure of the public to shipment contents.

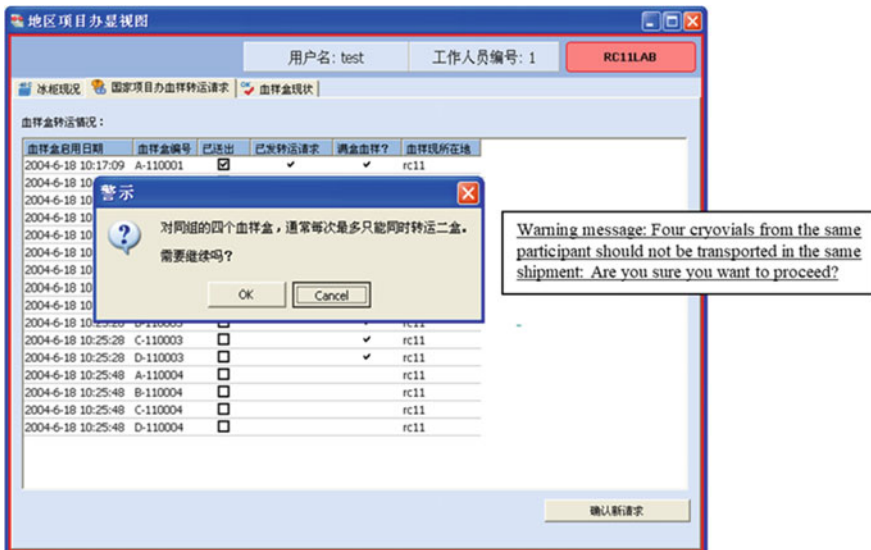


Fig. 4.5 Sample tracking software for managing sample shipment in CKB

## 4.6 Sample Storage and Retrieval

Biological samples collected from participants at baseline and other occasions are both invaluable and irreplaceable. For a large biobank, there may be millions of separate samples, so an essential aspect is to carefully plan both safe and secure storage and the ability to efficiently retrieve samples for analysis.

### 4.6.1 *Sample Storage*

Long-term storage of samples for large-scale biobanks is most efficient when using central facilities at a single, or few, locations. Storage facilities should have suitable equipment for the sample requirements, e.g., refrigerator or cold room (4 °C), freezers (−20 °C, −40 °C, −80 °C), or liquid nitrogen tanks (−200 °C), including capacity for immediate storage of newly arrived shipments and to allow for maintenance and equipment failures. Temperature control is critical, and a monitored alarm system should be in place. Back-up generators should be available to take over in case of power failure, and alternative storage locations should be identified as a contingency plan, e.g., in case of natural disaster or major equipment failure. For instance, CKB maintains substantial surplus storage capacity at −40 °C, available for emergency use, into which samples can be transferred if necessary. Where possible, duplicate samples from a participant should be stored in different locations, both in different freezers and on different sites, so that a single failure will not destroy or place at risk all samples from a participant.

In planning analyses of samples, especially where these require transfer of samples to other locations, it is important to have a record of the precise location of each sample in storage, and to track any sample movements including between boxes or between freezers. In CKB, IT systems record the locations of all samples held at the central sample storage site, at the following levels: freezer or tank number; shelf number; rack number; box number; box position (see Fig. 4.6).

### 4.6.2 *Sample Retrieval*

Samples will need to be retrieved from storage for processing or analysis, and if the locations of samples in storage are precisely recorded and any movements monitored, then it should be straightforward to locate them in freezers or liquid nitrogen tanks and to then retrieve them. For large-scale procedures such as DNA extraction, samples may be processed box by box, but other analyses (e.g., nested case-control studies) may involve selecting individual samples. Where individual samples are being selected, to avoid damage or loss and to facilitate sample tracking (e.g., during shipment), they should be transferred to and stored in new boxes, with the sample

The screenshot displays the 'Main' window of the Womingshall IT system. At the top, it shows 'Womingshall Stats' with 'Current Kadostie box count at Womingshall: 11758' and 'Boxes with potential location conflict count: 0'. Search parameters include 'Study ID' (0), 'Cryovial Set ID', and 'Boxset Label' (KS-: 780001). A table below lists search results with columns: Tank, Rack, Slot, Aliquot, Box Number, Box Pos, Cryovial ID, Study ID, Date Arrived, Thaw Count, and Notes. Below the table is a 'Womingshall Unit 17 warehouse' layout showing a grid of storage boxes, with box 9 highlighted in red. A detailed view of box 9 shows a rack code 'A4', a cryovial ID '010041001', and various attributes like 'Aliquot A', 'Position 1', 'Lipaemic No', 'Haemolyzed No', and 'Empty No'.

Tank	Rack	Slot	Aliquot	Box Number	Box Pos	Cryovial ID	Study ID	Date Arrived	Thaw Count	Notes
9	A4	780001	A	780001	1	010041001	780000057	2005-06-30 08:...	N/A	
11	A4	780001	B	780001	1	010041001	780000057	2005-06-30 08:...	N/A	
9	A4	780001	A	780001	2	010041002	780000877	2005-06-30 08:...	N/A	
11	A4	780001	B	780001	2	010041002	780000877	2005-06-30 08:...	N/A	

Fig. 4.6 IT system recording the locations of samples in central CKB storage facilities

movements being recorded. Lists of samples should be organised to enable efficient sample retrieval and to avoid multiple opening of freezers or liquid nitrogen tanks, which can affect their temperature control. CKB developed IT systems for generating sample lists and tracking the movement of those samples to new storage boxes. While samples are being retrieved their temperature should be controlled, for instance, by placing the boxes on dry ice.

## 4.7 Strategies for Sample Analysis

Biological samples are a limited resource, and assays and analyses using them should be carefully planned to avoid wasting, degrading, or otherwise excessively depleting them. In addition to the constraints on the analyses that can be performed depending on the types and amounts of the samples collected, there are numerous other factors that need to be considered when planning sample analyses, including, for example: (1) some assays of interest may require samples that have not been

through multiple freeze-thaw cycles; (2) assay service providers may require samples formatted on micro-titre plates in a range of different formats; (3) the volume of sample required for different assays may vary widely, from a few  $\mu\text{L}$  to several mL; and (4) funding constraints may limit the number of samples that can be analysed. If future needs are not considered before the first round of analyses, particular options may no longer be available.

### 4.7.1 *Research Opportunities*

Historically, it has only been possible to measure a limited number of biomarkers in any one study, and/or a study's analysis strategy may have been directed by the original study objectives, perhaps focussing on a particular research question. Such approaches remain valuable, since they can often provide greater detail and depth than a more general approach. With recent advancements in analytical methods and technology, however, analyses can now be made of hundreds, thousands, or even millions of biological data points in a single assay, often using only a small portion of the available sample at high throughput. These types of "omics" analyses include whole genome sequencing or genotyping, DNA methylation arrays, measures of transcript abundance, and measurement of circulating proteins or metabolites (see Table 4.4). Other technologies (e.g., lipidomics, ionomics) can also provide large amounts of data per sample, but at somewhat lower throughput. Using multi-omics assay strategies can maximise the data generated, using a minimal amount of sample in a cost-effective and efficient way, enabling parallel research into a wide range of diseases and risk factors.

It will often be desirable to conduct trial analyses, perhaps in a small number of samples, for instance, to explore the value of type of measurement for a particular research question, or to investigate a new or developing technology. It is useful to have available a collection of samples that are otherwise surplus to requirements, which can be used for such exploratory assays rather than depleting valuable baseline samples. Options include the use of duplicate samples collected at the time of recruitment, either by design or by accident (in CKB, around 3000 participants went through the entire baseline procedure on two separate occasions), identified from amongst samples whose linkage to participants is in doubt (e.g.,

**Table 4.4** Omics analysis strategies in large biobank studies

Type	Analysis	Numbers of biomarkers
Genomics	Genotyping, sequencing	20 million–3 billion
Epigenomics	DNA methylation array, Bisulphite sequencing	800,000–3.3 million
Transcriptomics	RNA array, RNaseq	20,000–60,000
Proteomics	Proteins	>4000
Metabolomics	Metabolites	>1000

due to sampling handling errors), or collected on a separate occasion according to the same sample handling and storage procedures.

### **4.7.2 Analysis Study Designs**

Sample use for analyses needs to be carefully planned. It is preferable to measure analytes in all participants in the whole biobank study. In addition to generating the maximum amount of information, and enabling a wide range of research questions to be investigated, this is important in maintaining consistent analysis procedures, minimising depletion of the samples, limiting variation between different batches of samples analysed at different times, and avoiding the generation of biases and/or reducing the statistical power of subsequent analyses (e.g., due to over-representation of participants with particular disease endpoints of interest). However, for reasons of cost and resources (particularly for some of the newer 'omics technologies) it may only be possible to assay a subset of participants, or to conduct assays in stages. In this case, nested case-control studies within the biobank may be an option, selecting cases of particular disease(s) and a set of matched or cohort-based controls for measurements (see Chap. 1). However, this approach can be complex and time-consuming to set up, and may complicate or preclude comparison with measurements made in other subsets of the study (Conroy et al. 2019). Sample depletion must be carefully considered, and freeze-thaw cycles which can damage samples should be minimised.

A particular issue with assaying subsets of the cohort in stages is that batch effects can occur, with assay results affected by non-biological processes, such as differences in equipment, reagents, or processing methods. To avoid batch effects, samples should be assayed in random order, particularly if sets of cases and controls are being assayed. Other points to consider in planning sample analyses include replicate measurements of (a proportion of) samples, to assess the technical reproducibility of results, and assaying repeat samples from the same participant at different time points, to measure the extent of biological changes in biomarkers. The latter repeat measurements are particularly important in assessing the strength of any relationship between a measured biomarker and subsequent disease.

## **4.8 Monitoring and Troubleshooting**

Errors in sample handling rarely come to light at the time they occur, frequently only emerging when biomarker assays give results which are unlikely or impossible. Common examples include: (1) mismatches between a participant's reported gender and their sex as inferred from genetic data, or from measurements of hormones (e.g., testosterone); (2) values for key metabolites which are outside the expected range (e.g., very low levels of blood glucose incompatible with consciousness); and

(3) measurements do not match a participant's characteristics (e.g., high nicotine metabolite levels in a reported non-smoker). Individual mismatches do not necessarily reflect sample handling errors, but clusters of similar errors may be indicative of an error such as linkage errors due to a box of cryovials being rotated, swapped, or mislabelled during sample extraction or sub-aliquoting.

If such errors were not corrected, this would result in it being necessary to discard all data, not only for that assay in the mismatching samples but for all samples and all assays potentially affected. This is where scrupulous and detailed tracking of sample location history, linkages, and sample handling events are invaluable, and why it is important to monitor these rigorously through primary identifiers (numbers and barcodes) on tubes and boxes, whether attached on stickers (e.g., at time of sample collection) or as part of the manufacturing process. Typically, if sufficiently detailed records have been kept then the pattern of mismatches within blocks of linkage errors will enable identification of the error that led to the mismatches and correction of the linkage errors. In CKB, for example, sex mismatches in genotyping data revealed over 1000 linkage errors (1% of samples), which could be unambiguously corrected by identification of the precise sample handling error, during sample collection, DNA extraction, or sub-aliquoting.

Sufficient recording of sample handling can also enable correction of other errors, such as occurred when a subset of UK Biobank blood plasma samples underwent inadvertent dilution during sub-aliquoting (UK Biobank 2019). Downstream clinical biomarker assays in the affected samples could only be adjusted because of the detailed records of which samples were aliquoted at the time these errors occurred.

## 4.9 Summary

Biological samples from participants are one of the most important resources available to a biobank study. Every aspect of sample collection, processing, storage, retrieval, and analysis needs to be carefully planned in advance, with particular attention to ensuring long-term sample integrity. Of paramount importance is maintaining robust sample records and tracking, to ensure correct linkage between the results of sample analyses and other data originating from the participant who provided the sample. To fulfil these goals, it is essential to develop clear protocols, detailed SOPs, and robust IT systems, and to ensure that they are rigorously implemented, as was done for CKB. Such careful attention to the collection and use of biological samples has the potential to transform a biobank study.

## References

- Chen Z, Chen J, Collins R, Guo Y, Peto R, Wu F, et al. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int J Epidemiol.* 2011;40:1652–66.
- Conroy M, Sellors J, Effingham M, Littlejohns TJ, Boultonwood C, Gillions L, et al. The advantages of UK Biobank's open-access strategy for health research. *J Intern Med.* 2019;286:389–97.
- Elliott P, Peakman TC, Biobank UK. 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.
- Patil S, Majumdar B, Awan KH, Sarode GS, Sarode SC, Gadbaile AR, et al. Cancer oriented biobanks: a comprehensive review. *Oncol Rev.* 2018;12:357.
- 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, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* 2015;12:e1001779.
- UK Biobank. Biomarker assay quality procedures: approaches used to minimise systematic and random errors (and the wider epidemiological implications). 2019. [http://biobank.ndph.ox.ac.uk/showcase/showcase/docs/biomarker\\_issues.pdf](http://biobank.ndph.ox.ac.uk/showcase/showcase/docs/biomarker_issues.pdf)

# Chapter 5

## Monitoring Long-Term Health Outcomes of Biobank Participants by Record Linkages



Ling Yang and Zhengming Chen

### Contents

5.1 Introduction .....	100
5.2 Scope of the Work .....	101
5.3 General Approaches .....	103
5.4 Establishing Linkages for Follow-Up .....	107
5.5 Obtaining Follow-Up Data from Multiple Sources .....	110
5.6 Coding and Integrating Health Outcomes .....	116
5.7 Monitoring and Managing Outcome Data .....	118
5.8 Summary .....	120
References .....	121

**Abstract** Large prospective biobank studies can provide reliable assessment of the relevance of lifestyle, environmental and genetic factors, and their complex interplay, for disease aetiology. The value of prospective studies depends not only on the recruitment of a large number of individuals but also on the ability to monitor the health outcomes of participants over time. In large prospective studies of relatively healthy adults, only a small proportion of the participants will die or develop any particular diseases each year, so the follow-up needs to continue for a prolonged time period. As well as follow-up for cause-specific mortality, follow-up for non-fatal disease outcomes is extremely important, and will greatly increase the power and range of diseases that can be investigated. There are major challenges involved in maintaining the study cohort long-term and managing the reliable collection and ascertainment of high-quality outcome data for many different conditions. This chapter provides a general overview of the scope of work and practical procedures relevant for long-term follow-up for fatal and non-fatal health outcomes in large biobank studies.

---

L. Yang · 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](mailto:zhengming.chen@ndph.ox.ac.uk)

© Springer Nature Singapore Pte Ltd. 2020

Z. Chen (ed.), *Population Biobank Studies: A Practical Guide*,

[https://doi.org/10.1007/978-981-15-7666-9\\_5](https://doi.org/10.1007/978-981-15-7666-9_5)



**Keywords** Prospective studies · Biobanks · Follow-up · Health outcomes · Mortality · Morbidity · ICD

## Abbreviations

DOB	Data of birth
DSP	Disease Surveillance Points
HI	Health Insurance
HID	Health insurance number
IARC	International Agency for Research on Cancer
ICD	International classification of diseases
IT	Information technology
NCD	Non-communicable chronic diseases
NID	National Identity Number
VA	Verbal autopsy
YOB	Year of birth

## 5.1 Introduction

Large prospective biobank studies can provide reliable assessment of the relevance of lifestyle, environmental and genetic factors, and their complex interplay, for disease aetiology. The value of prospective studies depends not only on the recruitment of a large number of individuals with extensive collection of exposure data but also on the ability to monitor the health of participants over time. Since prospective studies usually recruit relatively healthy adults from general populations, only a small proportion of the participants will die or develop any particular diseases each year. So, the follow-up needs to continue for many years or decades in order to accrue large number of cases with particular conditions. Apart from cause-specific mortality, follow-up for non-fatal disease outcomes should also be prioritised, which, if feasible, will greatly increase the study power and the range of diseases that can be studied. The key requirement for long-term follow-up is to ensure that data for a wide range of health outcomes occurring among study participants are collected completely, consistently, accurately, and in a timely manner. Moreover, the study cohort should be properly maintained to minimise the loss of follow-up over time. To achieve these goals, careful planning and development of efficient and cost-effective methods for long-term follow-up are essential. This chapter provides a general overview of scope of work related to the long-term follow-up and possible data sources that may be considered in prospective studies. Moreover, it describes practical procedures and systems necessary for undertaking the required work at scale and in a timely manner.

## 5.2 Scope of the Work

Depending on the study design and research objectives, a wide range of health-related outcomes can be considered in prospective studies. These may include cause-specific mortality and morbidity (e.g., cancer incidence), any episodes of hospitalisation, primary health care, and other health-related data. The methods of follow-up may vary according to the outcomes of interest, study settings, and established local or national health infrastructures. Different conditions may be captured by different information systems, or, the same disease events in particular individuals may be captured by different systems, facilitating cross-checks to improve validity, reliability, and consistency.

### 5.2.1 *Types of Health Outcomes*

In prospective studies of common non-communicable chronic diseases (NCDs), cause-specific mortality is the most commonly collected health outcome, as comprehensive registration of causes of death is generally available. While it represents one of the most important outcomes in human health, mortality data may not fully capture the natural history of certain diseases, especially those with slow and long developmental processes (e.g., COPD), or allow research into the aetiology of non-fatal conditions (e.g., eye diseases, bone disorders). Moreover, mortality data may suffer disproportionately, compared with disease incidence, from reverse causality biases when assessing the relevance of certain risk exposures (e.g., adiposity, levels of plasma lipids) for certain conditions (see Chap. 1). In prospective studies, collection of other health outcome data over and above cause-specific mortality will not only improve the study power and accuracy of disease diagnosis but also greatly increase the range of diseases that can be studied. Moreover, it may also allow different types of research to be undertaken, such as natural history and management of specific diseases (Chen et al. 2020). Table 5.1 summarises major types of health outcomes and their likely data sources that should be considered in prospective studies. Irrespective of the types of health outcomes and their data sources, attention should also be paid to verification and further characterisation of disease events collected (see Chap. 6).

### 5.2.2 *Main Data Sources*

In prospective studies, the data sources and systems for obtaining health-related information may vary depending on local infrastructures, health care systems, and necessary permissions that may need to be secured (Table 5.1). In general, the study would not be considered appropriate in populations where it is not possible to collect

**Table 5.1** Main types and likely data sources of health outcomes in prospective studies

Outcome	Likely data sources
Cause-specific mortality	<ul style="list-style-type: none"> <li>• Death registry</li> <li>• Special surveys</li> </ul>
Disease morbidity	<ul style="list-style-type: none"> <li>• Cancer registry</li> <li>• CVD registry</li> <li>• Special surveys</li> </ul>
Hospitalization records	<ul style="list-style-type: none"> <li>• Hospital Episode Statistics (HES)</li> <li>• Health insurance claim databases</li> <li>• Special surveys and self-reported records</li> </ul>
Other health outcomes	<ul style="list-style-type: none"> <li>• Primary health care records</li> <li>• Dental and occupational records</li> <li>• Cancer screening databases</li> <li>• Mental health databases</li> <li>• Children health records and educational results</li> </ul>

reliable information on cause-specific mortality, which should be a minimal requirement for large prospective studies focusing on major NCDs. Where possible, information on death should be collected through official death registries, which provide detailed information about the cause of deaths recorded in the official death certificates. Death certificates are essential legal documents required in many, if not all, countries when people die. Death certificates are usually issued either by a medical practitioner certifying the deceased status of an individual or by an official registrar according to the date, location, and likely causes of death reported by family members of deceased individuals. Death registries tend to cover whole or large proportions of well-defined populations in particular regions. With an established track record, they have long been used in a wide range of epidemiological studies and should be first prioritised in large prospective studies.

Disease registries provide information on occurrences of a specific disease or disease category in general or targeted populations. The most common disease registry is cancer registry, although increasingly similar registries have become available for several other major diseases (e.g., stroke, ischaemic heart diseases) in certain countries. Compared with death registries, disease registries often include more detailed clinical information including disease diagnostic procedures (e.g., more detailed cancer histology, laboratory tests) but their coverage, quality, and completeness of the information collected may vary. Cancer registries are well established in many countries and generally follow the standard procedures set by the International Agency for Research on Cancer (IARC). Other specific disease registries tend to follow certain local health administration requirements. As most of these registries are often established without strict legal requirements and there may be more technique challenges, and the coverage and quality of information collected may be less optimal compared with death and cancer registries.

In many countries it is also possible for studies to link to a range of other health information systems. Specific examples include Hospital Episodes Statistics (HES) in the UK (Herbert et al. 2017), health insurance (HI) claim databases in China (Levy

et al. 2020), and primary health care records in the UK and many Nordic countries (Sudlow et al. 2015). Some of these systems (e.g., HES, HI) are routinely used to monitor the use and costs of hospital services in general populations. Although they are not established for research purposes, they may still provide a wide range of relevant information, both inpatient and out-patient, about disease diagnosis and management for prospective studies. In some instances, however, their use in prospective studies may not be straightforward and requires a lot of careful planning and efforts. Apart from the need to seek formal approval from relevant government agencies for data access, there may be issues with data completeness, quality, and consistency. For example, HI systems in China do not generally provide detailed information about disease subtypes, because these are not directly relevant for reimbursement purposes (Chen et al. 2011). Moreover, the database structure, naming of variables and coding systems used for disease diagnoses and medical procedures may vary across different areas and also change over time, posing major challenges for data integration and standardisation.

### 5.2.3 *Minimising Loss to Follow-up*

Ideally, all participants in prospective studies should be followed up completely from the beginning (i.e., enrolment into the study at the initial baseline survey or equivalent) to the end of study (or death, whichever comes sooner). In reality, it may be impracticable to achieve full follow-up, as some people may move, change address, or even migrate abroad at certain time points during the course of the follow-up period. These individuals may be considered as lost to follow-up, if their vital or health status cannot be traced and ascertained reliably. Loss to follow-up represents a major challenge in prospective studies, for it will not only reduce the study power but also introduce major biases in analyses, if the rates of loss to follow-up are high and differ importantly between exposure and non-exposure groups of interest (see Chap. 1). Hence, irrespective of the study design and types of outcome collected, great attention should be paid to regular updates of participants' contact details or other relevant information to help minimise loss to follow-up. Such information may be sought, subject to prior consent and approval, *passively* through linkage with residential records (or equivalent), or *actively* through periodic contacts by home visits, post, telephone calls, and emails with participants or their family members or relatives.

## 5.3 General Approaches

Depending on the study objectives, available resources, outcome measures, and local infrastructure, different follow-up approaches can be employed in prospective studies. These should be carefully planned and piloted before starting the project to help

**Table 5.2** Advantages and disadvantages of passive versus active follow-up

Methods	Advantages	Disadvantages
Passive follow-up	<ul style="list-style-type: none"> <li>• Cost-effective</li> <li>• Highly efficient</li> <li>• Disease diagnoses tend to be reliable</li> <li>• Data collection can be timely and more regular</li> <li>• Coverage tends to be complete</li> <li>• Less bias even with incomplete coverage</li> </ul>	<ul style="list-style-type: none"> <li>• Prior consent by participants is needed</li> <li>• Permission from agencies for access may be problematic</li> <li>• Coverage and data quality may vary and change over time</li> <li>• Linkages with source data may not be accurate</li> </ul>
Active follow-up	<ul style="list-style-type: none"> <li>• Enables collection of health outcomes not captured in health care</li> <li>• Can be combined with repeated assessment of risk exposures</li> <li>• No need to seek specific approval from government agencies</li> <li>• Good reliability in matching of participants over time</li> </ul>	<ul style="list-style-type: none"> <li>• Low response rate</li> <li>• Relatively poor reliability in reported disease diagnoses</li> <li>• High under-reporting rates in disease occurrence</li> <li>• Costly and time consuming</li> <li>• Poor long-term sustainability</li> <li>• Biases when response rate is low</li> </ul>

assess their feasibility, data quality and completeness, long-term sustainability, and resource implications. In principle, there are two totally different approaches for follow-up of health outcomes, one is achieved through direct contact with and/or report by study participants or their relatives (i.e., “active” follow-up) and the other one is relied on linkages with available official registries or health-related systems (i.e., “passive” follow-up) without direct involvement of study participants. While each has its advantages and disadvantages (Table 5.2), both approaches may be used simultaneously in a single prospective study.

### 5.3.1 *Passive Follow-Up*

In large prospective studies, passive follow-up represents the most efficient, reliable, and cost-effective way of obtaining a wide range of health outcomes among study participants. It is generally achieved through linkages, via unique personal identification numbers and/or certain matching algorithms, with established registries and other health-related systems available in study areas, without the need to directly contact and engage participants.

To enable this, prior consents from participants at recruitment are needed in order to access their health records by study investigators. In many countries, there are established national/regional death and cancer registries that are managed by different government agencies, but prior permission for access will be required. As for a range of other health outcome data (e.g., hospitalisation episodes, primary health care data), since they generally contain far more personal and sensitive information compared with death and cancer registries, it may be less straightforward to obtain access permission.

**Table 5.3** Overview of long-term follow-up for health outcomes in CKB

Data sources	Outcomes captured	Disease coding	Reporting methods	Reporting frequency
Mortality registry	Cause-specific deaths	ICD-10	Electronic	Monthly
Disease registry	Cancer, stroke, IHD and diabetes	ICD-10	Electronic	Quarterly
HI claim system	Any hospitalised event	ICD-10	Semi-electronic	Biannually

Depending on the work plan and prior agreement with agencies who manage the relevant data systems, the passive follow-up process can be undertaken on a regular basis, e.g., annually or every 6 months. In most circumstances, it can be managed electronically through the established linkages (as illustrated in Table 5.3). In populations without widespread use of unique personal identification (ID) numbers, the matching and record linkages may be dependent critically on certain matching algorithms, which can lead to mismatching and/or under-reporting. While passive follow-up can allow reliable collection of major health outcomes, it will not allow collection of certain other health data that may not be captured effectively by health service providers such as symptoms, cognitive functions, psychological state, and long-term medication use. Moreover, the diagnostic criteria and coding methods may differ between agencies or health providers, or change over time, which could pose challenges for data consistency and standardisation. Furthermore, a small proportion of study participants may not be covered by disease registries, health services or health insurance schemes, and their health outcomes may need to be obtained through active follow-up.

### 5.3.2 Active Follow-Up

In countries or populations without properly established death and disease registries, an alternative approach for tracing health outcomes in prospective studies is through active follow-up. This is usually done through direct contact with the participants by mail, phone calls, internet, or by face-to-face interview at participant's home, study clinics, or assessment centres. In certain circumstances (e.g., death), it may have to involve family members of participants. Apart from major health outcomes (e.g., death, hospital admissions), active follow-up will also enable collection of a range of other health-related data that are not routinely captured by conventional mortality/morbidity registries, such as cognitive and physical functional state, mental and psychological profiles, and long-term use of medications. However, active follow-up has a number of important limitations, including (1) it is expensive and time consuming to organise as it requires participant re-engagement; (2) its coverage is usually incomplete (response rate typically <70%) and may differ across different exposure groups, resulting in major biases in analyses; (3) health outcome data

captured are less reliable, even in high-income countries, and may often need further verification against hospital records (Jacobs et al. 2017); and (4) it can only be done periodically, e.g., every few years or combined with repeated assessment of risk exposures at resurveys.

### 5.3.3 IT Support

To facilitate long-term follow-up, establishment of reliable IT systems is required to support all aspects of the work. The systems may vary in their functionalities, complexity, and types of platform used depending on the study settings, needs, and follow-up methods. Where feasible, they should be developed and maintained internally by the study IT team (see Chap. 7). Apart from collecting health outcome data (e.g., mortality and morbidity), such IT systems may also be used to collect unique personal identification numbers automatically and facilitate disease coding/standardisation, monitoring, and outcome verification and adjudication (see Chap. 6), as illustrated through examples in CKB (Table 5.4). The development and implementation of IT systems should be carefully planned and properly tested (see Chap. 7). Their use in large prospective studies will help ensure reliable,

**Table 5.4** Overview of IT systems related to long-term follow-up for health outcomes in CKB

Software name	Platform	Main functions	Users
LTFollow-up	Desktop	Data entering/viewing/editing program for participants' mortality and morbidity follow-up data and changes of contact information	Staff involved in long-term follow-up data collection
DrList	Desktop	Generates lists of participants' contact information in each clinical location for linkages and field work	Staff involved in active follow-up or surveys
NID card reader	Laptop	Portable device to scan/view/record information stored in the National ID cards	Regional staff involved in disease surveillance
PVD	Tablets	Collects hospital admission/clinical information from medical notes for selected conditions	Regional staff involved in disease surveillance
<i>i</i> -Case	Web	Allows specialists to undertake online event adjudication through reviewing medical records collected by PVD	Members of outcome adjudication committees
Standardiser	Desktop	Standardises and codes disease diagnosis reported through HI system	Clinical staff involved in code standardisation
Reporting	Desktop	Tool for summarising and monitoring mortality and morbidity data collected	Data managers/statisticians/study administrators
Teleport	Desktop	Tool for transferring collected long-term follow-up data from regional to central study offices	All staff involved in long-term follow-up data collection

complete, consistent, and timely collection of outcome data over time and across different study areas.

## 5.4 Establishing Linkages for Follow-Up

Prospective studies will take many years to complete the recruitment of large numbers of individuals and some study participants may die or develop disease events soon after joining the study. As such, follow-up for health outcomes among enrolled study participants should be initiated as soon as possible without waiting for the completion of the baseline survey. To manage the process effectively, development of relevant procedures and systems for follow-up should commence more or less at the same time as for the baseline survey, including obtaining relevant approvals for data access, development of IT systems, processing of personal information, and establishment of reliable record linkages with relevant external sources. Once the follow-up has started, it is also necessary to develop detailed plans and protocols for data integration, standardisation, monitoring and management (see Chap. 8).

### 5.4.1 *Obtaining Formal Approvals*

In passive follow-up, information about the health outcomes of participants are generally sought indirectly from external sources that are usually managed by government agencies. As well as obtaining informed consent from participants at the time of their enrolment (see Chap. 2), formal approvals from relevant agencies would also be required in order to access participants' health records. Even with participants' formal consent, there is no guarantee that the permission will be granted by these agencies, for protection of personal data has always been a politically sensitive and evolving issue. There are many international and national regulations governing data protection, which are issued by different organisations and may be subject to different interpretations, especially when the data are to be used outside of the initial scope under which they were first created (Staunton et al. 2019). The process for obtaining approvals may vary depending on the data sources, sensitivities of the information sought, and the perceived risks if released or leaked unintentionally. For certain outcome data (e.g., cause-specific mortality and cancer incidence) that have been widely used by researchers, it should generally be straightforward to obtain approvals from relevant agencies, while for other more detailed clinical information such as HI and primary care data, it may not be possible at all or may take many years of negotiations before formal, and often limited, approval can be granted. To help facilitate the approval process, it would be helpful to limit the data sought only to those that are needed for the planned research (e.g., information on disease diagnosis rather than costs related to disease management in hospital).



Once an approval has been granted, it is necessary to have a formal agreement with the relevant agencies to cover a range of issues, including the scope of data to be provided, time schedule, and associated costs. To foster long-term relationships, it would be helpful to keep data providers informed of study progress and findings, or even invite relevant individuals from particular agencies to attend regular collaborators meetings.

### ***5.4.2 Processing Personal Information***

At the initial enrolment into the study, various types of personal information should have been collected. Of these, a unique personal ID number would be the most important for long-term follow-up. In different countries, the personal ID numbers may vary in their official name, format, and scope of general use (e.g., NHS number in the UK, Social Security Number [SSN] in the USA, National Identity Number in China). These personal ID systems often have universal coverage, with a unique ID number usually allocated to an individual at birth. Moreover, they are widely used in various settings, including residence/household registry, health insurance, and health service, thus providing the most appropriate key for linking participants with external data sources.

In populations without reliable nationwide personal ID systems, use of other personal, albeit less unique, identifiers (e.g., name, sex, date of birth, race/ethnicity) may represent the only viable option to enable linkages with different data sources. To facilitate long-term follow-up, the information sought from participants at their initial enrolment should be as complete and comprehensive as possible, subject to their consent. For example, apart from standard contact information (i.e., address, phone number), information about personal email or social media accounts (e.g., WeChat, WhatsApp) should also be collected, along with their next of kin's contact information. Moreover, this information should be updated periodically, where possible, through review of official residential records or by active follow-up.

Where possible, the personal identifiers, particularly unique personal ID number, should not be entered manually (as they tend to contain long digits) but automatically using special device and IT systems at the initial recruitment phase in order to minimise data entry error (see Fig. 5.1). Alternatively, they may be obtained electronically from local residential records or other (e.g., electoral) registries, which can then be processed and properly integrated, after further verification by participants at enrolment, into study database. For linkage purposes, the personal identifiers can be processed, compiled, and released in batches according to communities/areas and time of enrolment, so that follow-up for outcome can start with as little delay as possible among participants who have already been enrolled into the study.



Fig. 5.1 Bespoke IT system to automatically capture personal ID information in CKB

### 5.4.3 Establishing Record Linkages

To help establish reliable linkages with external data sources, a standardised record linkage protocol should be developed and properly tested beforehand that includes matching methods and data linkage procedures which internal and external staff will follow during the process. Where necessary, proper training should be provided to individuals involved, along with instruction manuals for use at local study centres and/or relevant agencies.

In follow-up, precise matching of personal information is a prerequisite for ensuring reliable data linkages with external health information system. Depending on the study settings, types of the personal data available across different data sources, and the likely regulatory constraints, different methods may be used, ranging from a simple match of unique personal ID number, through to use of less unique personal information (e.g., name, DOB, sex), and to purely probability-based algorithms developed without full personal names. Although unique personal ID numbers would generally yield the most reliable matching for most individuals, their use may still not be perfect. For example, the personal ID number may be incorrect, or may change due to replacement for a lost one. To help reduce both false positive and negative matching rates, it is often necessary to consider using multiple matching methods, perhaps in a step-wise approach (Fig. 5.2).

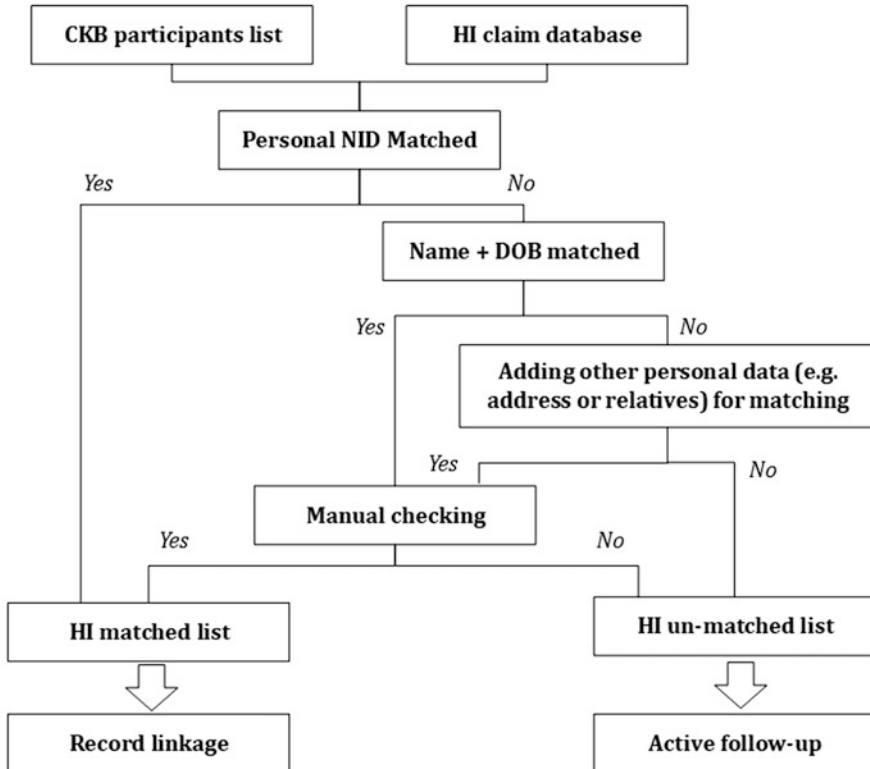


Fig. 5.2 Procedures for matching study participants with the HI system in CKB

Similarly, although much of the matching process would be done electronically, certain manual checks and review of different records would also be useful, especially for those with only partial matching. In some circumstances, it may not be possible to achieve a perfect match for certain individuals, for which a probability score should be generated that will enable study investigators to test and apply certain cut points to indicate a high likelihood of a true match. This may be particularly relevant for studies in which the data linkages are undertaken externally by agencies that hold the outcome data, without the opportunities for any manual review of partially matched records.

### 5.5 Obtaining Follow-Up Data from Multiple Sources

In most circumstances, a single data source may be insufficient to cover all cases or provide all necessary information about particular conditions. Linkages to multiple data sources would enable investigators not only to extend the range of outcomes that can be captured but also facilitate cross-checking and validation including

identifying previously undetected cases, even though it may greatly increase the workload and difficulties in data processing and integration. Once the linkages with external data sources have been properly established, the way with which to obtain the required health outcome data may vary depending on study settings and agreed data transfer protocols.

### ***5.5.1 Cause-Specific Mortality***

In most countries, the vital status of the general population is usually monitored through death registries which should be able to provide the study with all the cases of death occurring among the linked study participants. Most death registries tend to use international standard death certificates to record relevant information about each death, including contributing and underlying causes of death and their associated disease codes, along with diagnostic procedures (Fig. 5.3). With few exceptions, most countries now use ICD-10, the tenth revision of the International Statistical Classification of Diseases and Related Health Problems (ICD), which was introduced officially in 1994 to replace ICD-9. It is likely that ICD-11 may be introduced in the near future to replace ICD-10. While requesting cause-specific mortality data from official death registries, it is important to obtain not only ICD-10 codes of underlying and contributing causes of death but also all other relevant information captured on death certificates in order to undertake independent checks, including the appropriateness of the assigned underlying cause of death. Where feasible, it would also be useful to obtain a paper or electronically scanned copy of the official death certificate to further facilitate subsequent data checking and verification.

The frequency of data collection from death registries may depend on a number of factors (e.g., study sample size, prior agreement, and routine time delays in processing death certificates by death registries). In general, it would be appropriate to request cause-specific mortality data at least once a year. The mortality data may be sent to the study centre by agencies as an encrypted file, or using more direct data transfer methods (e.g., FTP), as is the case with CKB. In CKB, the study regional coordinating offices are based at local (rural county or urban district) offices of the Centre for Disease Control and Prevention (CDC) which is responsible for managing mortality registries (Yang et al. 2005). As such, the project was able to collect cause-specific mortality and other health outcome data more directly, frequently, and efficiently, using bespoke IT systems (see Fig. 5.3), which not only capture all information on death certificates but also enable certain automated logic checks of coding and assignment of underlying causes of death using specific algorithms.



### **5.5.2 Morbidity for Major Diseases**

In prospective studies the most commonly collected morbidity data are cancer incidence from cancer registries. In some countries cancer registries are managed by the same government agencies as those that manage death registries, while in others it may involve different agencies. In many LMICs with limited resources, cancer registries may only cover certain well-defined geographical regions rather than the whole country. Increasingly these regional registries may include a few other major diseases (e.g., stroke, MI), as is the case in China (Chen et al. 2011). As for cause-specific mortality, the information to be provided by disease registries should cover all essential data captured on the form and use similar data transfer strategies as described above. In CKB, all the ten study areas chosen have established local disease registries covering major categories of diseases (e.g., cancer, stroke, IHD, and diabetes). The occurrence of these diseases is usually notified and reported by local hospitals using standard procedures and forms, with varying degree of completeness and quality. Nevertheless, they do provide important additional information that is not captured on official death certificates, such as non-fatal events and more detailed clinical information (e.g., pathology subtype and clinical stage information for cancer) (Fig. 5.4).

### **5.5.3 Episodes of Hospitalisation**

In countries with universal health care systems, it may be possible to collect episodes of hospitalisation or even primary care data among study participants. In the UK, it is possible to use personal NHS number or other approaches to link to HES data, which is a data warehouse containing details of all inpatient and outpatients admissions, including clinical information about diagnosis and operation. The HES database is managed by UK NHS, which provides a monthly data extraction service to approved users and researchers, and covers a wide range of disease codes using standard disease diagnosis and procedure codes used routinely in the NHS. Through a formal application process, it can provide various studies with specified information on a regular basis, as is the case with UK Biobank and the Million Women Study (Green et al. 2019; Sudlow et al. 2015). In Nordic countries, similar systems are also available to support population health studies. In recent decades, many East Asian countries, including China, have also established similar systems that can be used to support large biobank studies. In CKB, ~98% of study participants have been linked successfully, using their unique personal ID number and other information, to the Health Insurance (HI) system, which records any episodes of hospital admission including disease description, ICD-10 code, and diagnostic and treatment procedures. The HI system is managed locally at city/county level but follows national common protocols. However, it lacks certain specific disease information relevant for biobank studies, such as histological types of cancer, for which separate verification and adjudication through review of medical notes may be needed (see Chap. 6).

Long-Term Follow-Up Wizard

### Enter Disease Incidence Report

Enter into the system the required details from the paper disease incidence report

**No disease incidence report image exists**

Code  National ID 123456740402123

#### Disease Incidence Report

Incident  Correction

Event report source  Hospital  Date of birth 1947  4

Sex  Male  Female  Tel. no.

**Date of incidence**     Estimated

Date of diagnosis

Date of death

Clinical  Pathology  ECG  X-ray  Ultrasound  CT/MRI  Endoscopy

Immunology  Biochemistry  Surgery  Bone marrow  Cytology  Unknown

Diagnostic criteria (choose one or more)

Reporting hospital  Department  Department type  Ward  Outpatient  Village doctor

Diagnosing hospital  Department  Department type  Ward  Outpatient  Village doctor

Date of report    Rep./completing person

Date of arrival at CDC    CDC person  Notes

**Disease Diagnosis**

IHD  Stroke  Diabetes  Cancer (Old)  Cancer

Original description

Primary site  Bronchus and Lung (C34.0-C34.9)

Primary sub-site  Bronchus or lung (C34.9)

Metastatic site  Main bronchus (C34.0)

Pathological sub-type  Upper lobe, lung (C34.1)

Middle lobe, lung (C34.2)

Lower lobe, lung (C34.3)

Overlapping lesion of lung (C34.8)

Bronchus or lung (C34.9)

Pathological test no.

TNM classification  T  N  M

Cell diff. status

ICD-10  C34.9 支气管或肺, 未特指

Rev. dis. name

Fig. 5.4 Bespoke IT system to collect major disease morbidity in CKB





## 5.6 Coding and Integrating Health Outcomes

To facilitate research, all health outcomes collected from various sources should be coded using the ICD system. The ICD is published by the WHO and is designed to promote international comparability in the collection, processing, classification, and presentation of health statistics and medical research. The ICD classification is revised periodically to account for newly emerging conditions and revised definitions. The current tenth revision (ICD-10), covering >150,000 primary codes, was formally introduced during the mid-1990s, and ICD-11 is scheduled to come into effect in 2022. Apart from ICD, other coding systems may also be used by various health information providers in many countries, mainly in clinical settings to capture more detailed information about clinical presentations and management. Often it is necessary for particular studies to review, standardise, and map those codes to the ICD coding system so that a single and streamlined disease coding system can be used for research purposes. Likewise, prospective studies with prolonged follow-up also need to develop systems to map different versions of ICD to ensure consistency in the definitions of major diseases.

### 5.6.1 *Cause-Specific Mortality*

All official death registries should provide standard ICD codes for underlying and contributing causes of each death as certified by medical doctors or health professionals. Mortality coding involves two components—the correct assignment of ICD codes to the conditions reported on death certificates, and the correct application of various coding rules when determining the ‘underlying cause of death’, which is the disease or injury that initiated the sequence of events leading directly to death. Even though these codes are directly supplied to researchers by official registries, it is still necessary to check their validity independently, for the coding rules are complex and there may well be certain local practices that could systematically misinterpret certain rules. For example, in certain countries, individuals with diabetes who have subsequently died from heart attack may mistakenly have diabetes (ICD-10, E10-E14), rather than myocardial infarction (ICD-10, I21), assigned as the underlying cause of death. Without regular checking, such issues may be discovered only during data analyses, and a systematic review and re-coding of all the related deaths may be needed, as has happened in a large prospective biobank study in Mexico (Alegre-Diaz et al. 2016). To help minimise such errors many death registries nowadays have applied automated algorithms to check the accuracy and consistency of ICD coding for each death (NCHS 2017). Such algorithm can also be used by investigators in prospective studies as part of the routine data management activities.

## 5.6.2 *Cancer and Major Disease Incidence*

For malignant neoplasms, the International Classification of Diseases for Oncology (ICD-O) system has been developed to code cancer diagnoses and has been used for nearly 50 years by various cancer registries. ICD-O is a domain-specific extension of ICD for cancer and will provide specific codes for the site (topography) and the histology (morphology) of the neoplasm, usually obtained from a pathology report (Fritz et al. 2013). A checking and conversion programme has also been developed by IARC/WHO with mapping tables between ICD-10 and ICD-O systems. Apart from cancer, other disease registries (e.g., stroke, IHD) will probably supply ICD-10 codes, as in death registries.

## 5.6.3 *Episodes of Hospital Admissions*

The coding for episodes of hospital admission may involve ICD or other specific code systems developed by health care providers in each country. In the UK HES data warehouse, clinical diagnoses including co-morbidities are coded according to the ICD system, while all operations and procedures are coded according to the system developed by OPCS (Office of Population, Censuses and Surveys) (NHS 2019). Apart from ICD codes, many countries have developed their own, albeit more detailed and comprehensive, clinical coding systems, such as the READ Codes in the UK NHS, which is a comprehensive computerised coding system for clinicians to record all clinical terms and procedures in hospital and primary care settings. It also contains mapping tables which can be used to generate ICD-10 codes. The READ Codes system has been used for many decades and has had a number of updates. It is anticipated that it will be replaced by a new system—SNOWMED CT, which will have more detailed information about diagnosis and procedures, symptoms, family history, allergies, assessment tools, observations, devices to support clinical decision making (NHS 2019).

Depending on the study need and access agreement reached, the biobank studies may be supplied by agencies with standard ICD-10 codes, with or without the associated disease names. On the other hand, certain studies may be able to obtain more detailed clinical information for the study participants, as in the UK Biobank studies of READ codes captured in primary care (Sudlow et al. 2015). When supplied, the outcome data may not be in a readily usable format, for which certain processing work would be needed. As has happened in the CKB, the outcome data extracted from the HI system sometimes include unstandardized disease names, unrecognised disease codes, and multiple disease names, descriptions or ICD-10 codes joined together. To improve the data quality, a bespoke IT system has been developed in CKB to automatically split multiple diseases input in the same row, and then automatically map ICD-10 codes to each disease name (Fig. 5.5).

description	icd10_code1	icd10_code2	translation
7519 ● 冠心病 心肌梗塞 脑梗塞 慢性肝炎	I10		Coronary heart disease angina infarction, chronic hepatitis
7520 ● 冠心病	I25.1		Coronary heart disease
7521 ● 心绞痛	I20.9		Coronary heart disease angina
7522 ● 脑梗塞	I63.9		Cerebral infarction
7523 ● 慢性肝炎	K73.9		Chronic hepatitis
7524 ● 冠心病 心律失常 脑梗塞			
7525 ● 冠心病	I25.1		Coronary heart disease
7526 ● 心律失常	I49.9		Arrhythmia
7527 ● 脑梗塞	I63.9		Cerebral infarction
7528 ● 冠心病 高血压 脑梗塞 前列腺增生	I10		Coronary heart disease, hypertension, Cerebral infarction, Benign prostatic hyperplasia
7529 ● 冠心病	I25.1		Coronary heart disease
7530 ● 高血压	I10		hypertension
7531 ● 脑梗塞	I63.9		Cerebral infarction
7532 ● 前列腺增生	N40		Benign prostatic hyperplasia
7533 ● 冠心病 脑梗塞 不足	I25.1		Coronary heart disease, Cerebral insufficiency
7534 ● 冠心病	I25.1		Coronary heart disease
7535 ● 脑梗塞 不足	I63.8		Cerebral insufficiency
7536 ● 冠心病 糖尿病 脑梗塞	I10		
7537 ● 冠心病	I25.1		Coronary heart disease
7538 ● 糖尿病	E14.9		diabetes
7539 ● 脑梗塞	I63.9		Cerebral infarction

Fig. 5.5 Bespoke IT system for automated disease standardization in CKB

### 5.6.4 Integrating Outcomes from Different Sources

Once collected, the health outcome data should be checked and, if necessary, converted into standard ICD-10 codes or more streamlined disease endpoint codes to facilitate analyses (see Chap. 8). The data collected from different sources can be linked and integrated via the unique study ID numbers that are usually linked with personal identifiers (e.g., unique personal ID number and name of participants) supplied to data agencies. In large studies with many outcomes collected from different sources using different variables and data structures, it would be preferable to keep them separately in different databases rather than having them fully integrated into a large and complex single database. Depending on the future research need, certain key data such as ICD-10 codes for all or certain major diseases may be extracted to create a separate disease endpoint database. Alternatively, it may create a separate data file listing all outcomes captured from different sources for particular participants. These outcome data files can then be used in various way by researchers, including data checking to identify overlap and inconsistency such as an event reported after death (see Table 5.6).

## 5.7 Monitoring and Managing Outcome Data

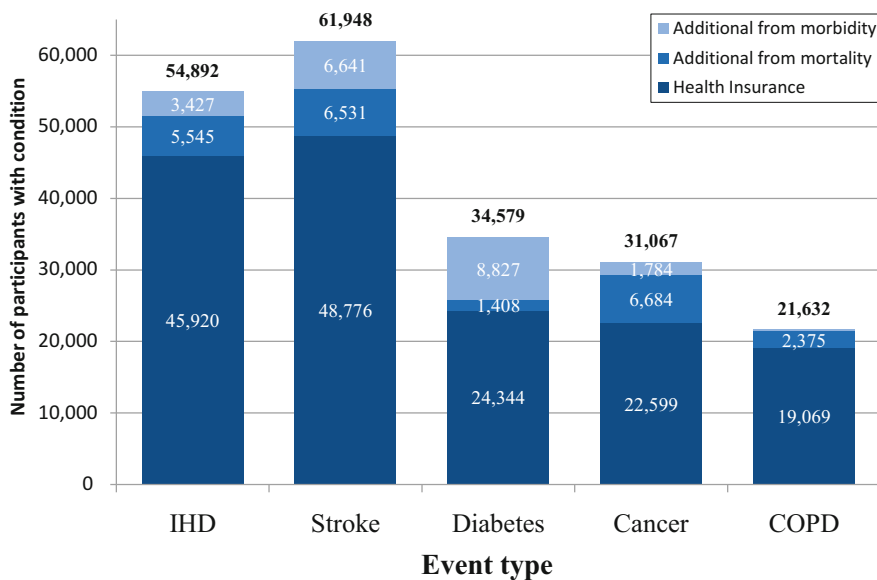
Although the external outcome datasets would be structured and coded, they are often set up for administrative purposes in real world settings, so inevitably the data quality may not be optimal. Detailed data quality monitoring and management plans should be developed and implemented at the early stages of the follow-up to ensure that health outcome data are collected reliably, consistently, completely, and in a

**Table 5.6** Example of all outcomes for one single participant from multiple sources in CKB

Participant ID	Source	Source variable	Diagnosis date	ICD-10 code (disease name)
990000811	Disease reporting	Stroke report	02-Jan-2009	I63 (ischaemic stroke)
990000811	Health insurance	Admission diagnosis	05-Jan-2009	I63 (ischaemic stroke)
990000811	Health insurance	Discharge diagnosis	07-Jan-2009	I61.1 (haemorrhagic stroke, cortical)
990000811	Disease reporting	Stroke report	18-Mar-2010	I61 (haemorrhagic stroke)
990000811	Death registry	Underlying cause	01-Jun-2010	I21 (heart attack)

timely manner. Initially, the main focus would be on the completeness, consistency, and quality of the data collected for particular individuals from particular sources. For example, for each reported death there should be basic checks for completeness of data for key items on the death certificate, internal consistency (e.g., age at death vs DOB and date of death), and appropriateness of assigned ICD-10 codes for underlying and contributing causes. As more outcome data become available, it would be necessary to undertake certain statistical monitoring to examine data consistency, general patterns, and timeliness of the data reported over time and across different study areas/centres. For deaths, these checks may include: (1) proportion of deaths with unknown and/or ill-defined causes; (2) proportion of deaths without certified contributing causes; (3) proportion of deaths occurring outside of hospital; (4) mean time delay between deaths and reporting; (5) proportion of deaths due to certain specific diseases (e.g., diabetes, hypertension) prone to coding errors; (6) diagnostic criteria for certain diseases (e.g., pathology for cancer); and (7) change of overall mortality rates over time and comparison with the general population. It should be noted that the key aim of the statistical monitoring is to detect any irregularities in reporting, rather than to improve external systems, so that any data issues revealed in the study are properly understood and so may inform subsequent analyses (e.g., re-coding of all reported cases for particular conditions).

Apart from an increased range of outcomes captured, combining data obtained from different sources can also greatly improve the completeness and quality of outcome data in the study. However, the process may not be straightforward and requires careful planning and development of appropriate procedures (see Chap. 8). Even after cleaning and standardisation, the outcome data from multiple sources for particular participants may cover different periods and contain different variables, duplicate events, or even conflicting information (e.g., a hospital admission after a death). For any conflicting information, it may not be possible to resolve any disparities across multiple sources for particular events. When absolutely necessary such problems can be resolved by introducing a hierarchy of trust, removing data from less reliable sources. It should be emphasised that during the process no data should be deleted permanently from the datasets. The main aim is to check, manage,



**Fig. 5.6** No. of incident cases of five major diseases from difference sources in CKB

and transform the outcome data from these disparate sources into something that can be readily used for analyses. To facilitate research, it may be necessary to generate simplified but integrated outcome datasets, containing: (1) one row per participant (using participant ID); (2) date of diagnosis; and (3) presence or absence of particular disease. Figure 5.6 shows the numbers of incident cases of five major diseases captured from different sources in CKB. It demonstrates the value of record linkages with multiple sources, along with the hierarchy of their data quality (with HI being the highest, followed by mortality and morbidity registries) relevant for data integration and analyses.

## 5.8 Summary

This chapter provides an overview of the general approaches and practical procedures involved in establishing and managing long-term follow-up in large prospective studies. To ensure that good quality outcome data can be obtained reliably, consistently, completely, and in a timely manner over a prolonged period, passive follow-up through record linkages with external data sources is essential. Nevertheless, active follow-up through direct contacts with participants and/or their relatives can be harnessed in various ways to serve the study needs. To establish reliable linkages, the collection of unique personal ID numbers and/or development of reliable matching algorithms based on less stringent personal identifiers are critical.

Moreover, standardised procedures and robust IT systems are required to support and facilitate the outcome data collection, processing, integration, and monitoring. To meet future research needs, further efforts may also be needed to verify and adjudicate disease events for particular diseases in order to improve disease sub-phenotypes, which will be discussed specifically in the next chapter.

## References

- Alegre-Diaz J, Herrington W, Lopez-Cervantes M, Gnatiuc L, Ramirez R, Hill M, et al. Diabetes and cause-specific mortality in Mexico City. *N Engl J Med*. 2016;375:1961–71.
- Chen Z, Chen J, Collins R, Guo Y, Peto R, Wu F, et al. 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 Y, Wright N, Guo Y, Turnbull I, Kartsonaki C, Yang L, et al. Mortality and recurrent vascular events after first incident stroke: a 9-year community-based study of 0.5 million chinese adults. *Lancet Glob Health*. 2020;8:e580–90.
- Fritz A, Percy C, Jack A, Shanmugaratnam K, Sobin L, Parkin DM, et al., editors. International classification of diseases for oncolgoy - third edition first revision. Lyon: International Agency for Research on Cancer; 2013.
- Green J, Reeves GK, Floud S, Barnes I, Cairns BJ, Gathani T, et al. Cohort profile: the million women study. *Int J Epidemiol*. 2019;48:28–29e.
- Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data resource profile: hospital episode statistics admitted patient care (HES apc). *Int J Epidemiol*. 2017;46:1093–1093i.
- Jacobs EJ, Briggs PJ, Deka A, Newton CC, Ward KC, Kohler BA, et al. Follow-up of a large prospective cohort in the United States using linkage with multiple state cancer registries. *Am J Epidemiol*. 2017;186:876–84.
- Levy M, Chen Y, Clarke R, Bennett D, Tan Y, Guo Y, et al. Socioeconomic differences in health-care use and outcomes for stroke and ischaemic heart disease in China during 2009–16: a prospective cohort study of 0.5 million adults. *Lancet Glob Health*. 2020;8:e591–602.
- NCHS. National vital statistics systems. 2017. <https://www.cdc.gov/nchs/nvss/instruction-manuals.htm>. Accessed 12 May 2020.
- NHS. NHS digital - terminology and classifications. 2019. <https://digital.nhs.uk/services/terminology-and-classifications>. Accessed 11 Apr 2020.
- Staunton C, Slokenberga S, Mascalconi D. The gdpr and the research exemption: considerations on the necessary safeguards for research biobanks. *Eur J Hum Genet*. 2019;27:1159–67.
- Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12:e1001779.
- WHO. Verbal autopsy standards: ascertaining and attributing cause of death. 2007. <https://apps.who.int/iris/handle/10665/43764>
- Yang G, Hu J, Rao KQ, Ma J, Rao C, Lopez AD. Mortality registration and surveillance in China: history, current situation and challenges. *Popul Health Metrics*. 2005;3:3.

# Chapter 6

## Verification and Adjudication of Health Outcomes in Prospective Cohort Studies



Yiping Chen and Robert Clarke

### Contents

6.1 Introduction .....	125
6.2 Guiding Principles and Approaches .....	126
6.3 Practical Procedures for Verifying Reported Outcomes .....	131
6.4 Practical Procedures for Outcome Adjudication .....	136
6.5 Monitoring and Data Management .....	142
6.6 Summary .....	142
References .....	143

**Abstract** The value of prospective biobank studies critically depends on their ability to collect large number of well-characterised disease outcomes on study participants over a prolonged period of time. In contrast with case-control studies which typically collect data on disease cases directly from hospitals, prospective studies collect incident disease outcomes during follow-up over several years or decades by linkage to death or disease registries. Verification of reported disease outcomes in prospective studies is the process of independent validation of the reporting sources for disease outcomes. Adjudication is the process of independent review of all the available evidence on clinical symptoms, signs, imaging, biochemical or histological investigations to classify reported disease outcomes into major disease types and their pathological and/or aetiological sub-types. Hence, disease verification and disease adjudication are complimentary processes to verify the accuracy of reported diagnoses and to classify major diseases into their pathological sub-types. Since most major diseases present as clinical syndromes, reliable classification of disease outcomes is required for studies of genetic and other determinants of such diseases. Verification and adjudication systems require collection of additional information from external sources, including disease registers or medical records from hospitals or primary health care systems for independent review by

---

Y. Chen (✉) · R. Clarke

Big Data Institute Building, Nuffield Department of Population Health, Old Road Campus,  
University of Oxford, Oxford, UK

e-mail: [yiping.chen@ndph.ox.ac.uk](mailto:yiping.chen@ndph.ox.ac.uk)

© Springer Nature Singapore Pte Ltd. 2020

Z. Chen (ed.), *Population Biobank Studies: A Practical Guide*,

[https://doi.org/10.1007/978-981-15-7666-9\\_6](https://doi.org/10.1007/978-981-15-7666-9_6)

123

clinical specialists. The design and implementation of practical procedures needed for disease verification and adjudication in large prospective studies requires feasible and cost-effective systems. Both verification and adjudication systems require regulatory approval to safeguard the confidentiality of personally identifiable data. This chapter discusses the principles and practical procedures required to establish systems for verification and adjudication of disease outcomes in large prospective studies, which will also be of general relevance for other studies.

**Keywords** Prospective studies · Biobanks · Health records · Adjudication · Disease classification

## Abbreviations

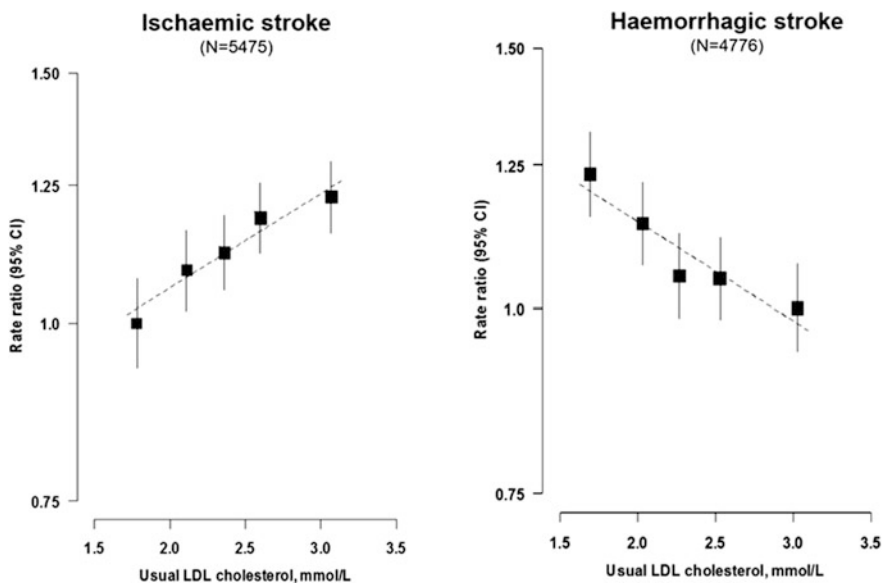
AI	Artificial intelligence
BMI	Body mass index
CDC	Centre for Disease Control
CKB	China Kadoorie Biobank
CK-MB	Isoenzyme of creatine kinase
COPD	Chronic obstructive pulmonary disease
CPRD	Clinical practice research datalink
CRF	Case report form
CT	Computed tomography
EHR	Electronic health records
GWAS	Genome-wide association study
HES	Hospital episode statistics
HI	Health insurance
<i>i</i> -CASE	Internet-based Case Adjudication System for clinical Events
ICD-10	10th revision of international classification of diseases
ID	Identification number
IHD	Ischaemic heart disease
LAA	Large artery atherosclerosis
MRI	Magnetic resonance imaging
NHS	National Health Service
PVD	Portable validation device
SBP	Systolic blood pressure
SOP	Standard operating procedures
SVD	Small vessel disease
TOAST	Trial of Org 10172 in Acute Stroke Treatment
VTE	Venous thromboembolism
WHO	World Health Organization



## 6.1 Introduction

The value of large prospective biobank studies relies not only on the recruitment of a large number of well-characterised healthy individuals but also on the collection of exposures (using questionnaires, physical measurements, and assays of biological samples) and disease outcomes during follow-up over several decades. The associations of exposures with particular disease outcomes may differ, both qualitatively and quantitatively, by disease types and their main sub-types. For example, recent findings from the China Kadoorie Biobank (CKB), a prospective study of >512,000 adults in China, showed that the associations of blood pressure with stroke were much stronger for intracerebral haemorrhage (ICH) than for ischaemic stroke (IS) (Lacey et al. 2018). Moreover, the associations of LDL-C with stroke differed qualitatively by stroke types, being positive for IS and inverse for ICH (see Fig. 6.1) (Sun et al. 2019). Likewise, the associations of adiposity with different oesophageal cancer sub-types differed qualitatively, being positive for adenocarcinoma and inverse for squamous cell carcinoma (Smith et al. 2008). Improved characterisation of disease outcomes in epidemiological studies has already not only enhanced our understanding of the relevance of many established risk factors for major diseases but also helped to identify novel genetic and non-genetic causes for many diseases (Herrett et al. 2013; Ay et al. 2014).

In case-control studies, data on particular disease outcomes are usually collected directly from hospitals. By contrast, prospective studies typically identify and record



**Fig. 6.1** Associations of plasma LDL-cholesterol with different stroke types in CKB (re-use with permission from Sun et al. 2019)

disease outcomes by linkage with mortality and disease registries. Such outcome measures tend to lack important clinical and diagnostic information needed to reliably classify diseases into relevant pathological sub-types. Hence, in addition to ensuring completeness of long-term follow-up (see Chap. 5), appropriate strategies are also needed in prospective studies to verify the reporting accuracy of disease outcomes and to further classify major diseases into their pathologically aetiologically relevant sub-types. The procedures for verification and adjudication may involve collection of additional clinical information (e.g., presenting symptoms and signs, results of laboratory tests, imaging or other diagnostic investigations) from medical records stored in hospitals or primary care systems for independent review by clinical specialists. As large prospective studies typically involve diverse regions (or countries) with disease diagnoses provided by multiple different hospitals, there are many regulatory, logistic, and practical challenges for undertaking disease verification and adjudication in real-world settings. Careful planning and development of secure and reliable procedures and systems are therefore required to ensure that the work can be carried out efficiently, cost-effectively, and at scale. This chapter describes general principles and practical procedures for disease verification and adjudication in prospective studies, with illustrative examples from CKB (Chen et al. 2005, 2011). Moreover, it considers some of the regulatory issues that may be encountered, in addition to the requirements for coordination and management of field work and clinical specialists for outcome adjudication.

## 6.2 Guiding Principles and Approaches

In many populations with limited health system infrastructure, collection of data on health outcomes may be limited to cause-specific mortality. However, increasingly in many other populations, it is now possible to obtain a wide range of fatal and non-fatal disease outcomes by linkage with cancer registries, Hospital Episode Statistics (HES), health insurance claims databases, and primary care records, such as Clinical Practice Research Datalink (CPRD) or other systems. While there are substantial challenges in accessing, linking, and integrating data from such sources (see Chap. 5), these systems can provide much more detailed clinical data to enable researchers to classify disease sub-types and address a wider range of research questions beyond analyses of cause-specific mortality.

### 6.2.1 *Main Objectives*

The chief objectives of outcome verification and adjudication in large prospective studies are to: (1) improve the reliability of diagnosis of cases (i.e., minimise false positive rates); (2) enhance specificity (i.e., minimise misclassification of disease outcomes); and (3) collect additional clinical and diagnostic information to reliably

classify and sub-classify disease outcomes. A multistage framework is needed to achieve these objectives in an efficient and cost-effective manner.

### 6.2.2 Scope of Work

Verification of reported health outcomes and adjudication of a subset of the most important outcomes involve a series of distinct, albeit integrated, stages of work to improve the quality of health outcomes recorded in prospective studies. These stages include: (1) ascertainment of suspected cases of disease or clinical events through linkage with available registries and health record systems (see Chap. 5); (2) disease verification through internal checks and cross-referencing, with or without review of source documents; and (3) disease adjudication with review of source documents. Each stage has distinct objectives and involves different data sources and procedures (Table 6.1).

Depending on the study design, research objectives, local health infrastructure, specific consent from participants, and regulatory constraints, the detailed plan and procedures required to undertake disease verification and adjudication are likely to vary substantially in different settings and between different studies. Moreover, even in the same study, the work involved for different diseases may also vary greatly, for the quality of routine health care data will not be uniform for different diseases. In many studies, the main tasks may simply involve collection and ascertainment of disease outcome data from a few major sources (e.g., death registries or cancer registries) and then undertaking internal checking on linkage quality, completeness,

**Table 6.1** Framework for outcome verification and adjudication in large prospective studies

Stage	Procedures	Likely data sources
1. Ascertainment of suspected cases	<ul style="list-style-type: none"> <li>Linkage to health records via unique participant ID or matching algorithm</li> </ul>	<ul style="list-style-type: none"> <li>Death registries</li> <li>Cancer registries</li> <li>Hospital admission records</li> <li>Health insurance records</li> <li>Primary care records</li> <li>Pharmacy records</li> </ul>
2. Verification of reported cases	<ul style="list-style-type: none"> <li>Cross-checking of disease registries and health records</li> <li>Source verification</li> </ul>	<ul style="list-style-type: none"> <li>Electronic health record (EHR)</li> <li>Disease registers</li> <li>Medical records (paper or electronic)</li> </ul>
3. Adjudication and classification of cases	<ul style="list-style-type: none"> <li>Adjudication by reviewing medical records</li> <li>Classification using pre-specified criteria</li> </ul>	<ul style="list-style-type: none"> <li>Medical records</li> <li>Imaging and other examinations</li> <li>Laboratory test reports</li> </ul>

removal of duplicate records and assessing data quality. In studies with more follow-up, the ability to detect possible under-reporting or other inconsistencies (e.g., dates of death occurring before admission to hospitals) and other data quality issues will certainly be improved by verification procedures (see Chap. 5). For example, the validity of coding of major diseases by HES system in the UK can be compared with Primary Care records (e.g., CPRD) or cancer registries and where the agreement is high, no further data collection may be warranted for verification or adjudication of HES data (Wright et al. 2012; Green et al. 2019). Even if the agreement is high, some routine checking is still indicated to ensure data quality over time. For many identified issues, it may be possible to resolve issues without the need to go back to source documents (e.g., original medical records). Conversely, it may not be possible or feasible, for regulatory or logistic reasons, to check source documents (e.g., original medical records) in order to correct any major errors or inconsistencies. In such circumstances, certain roles need to be developed to handle and accommodate any likely data issues identified, using data hierarchy systems based on perceived data quality of different reporting sources (see Chap. 8).

In many countries such as the UK and Scandinavia with well-established health care systems and data reporting infrastructures, disease outcomes recorded in some routine registries or electronic health records may be sufficiently accurate to reliably classify major diseases for most purposes. Despite this, it is still prudent for prospective studies to conduct independent validation of the quality of reported disease outcomes (Herrett et al. 2013). Such disease validation should involve review of objective records from independent external sources in a random sample of the individuals with major diseases. The study findings should help to estimate the positive predictive value of the reported diagnoses for particular disease outcomes in the study. For example, the Million Women Study (MWS) in the UK undertook validation study of disease diagnoses reported by HES using the more detailed primary care records, involving a random sample of 1000 cases of each of stroke, ischaemic heart diseases (IHD), and venous thromboembolism (VTE). This study demonstrated almost 95% agreement for these major vascular disease outcomes, while for sub-types of stroke, the agreement was somewhat more variable (e.g., 86% for ischaemic stroke, 78% for intracerebral haemorrhage) (Wright et al. 2012).

Access to and use of external records can be challenging, as such systems were not designed for research purposes, so the information routinely collected in these systems may be limited, especially for undertaking detailed phenotyping of certain specific diseases (e.g., Large Artery Atherosclerosis [LAA] vs Small Vessel Disease [SVD] IS) using diagnostic algorithms (e.g., TOAST) (Adams Jr et al. 1993). For certain conditions such as cancer, even the reported diagnosis may be very reliable, additional data are typically required on cancer sites, tumour histology, tumour stage and grade, results of cancer biomarkers (e.g., oestrogen receptor status), which may not be routinely available from certain reporting sources. Hence, collection of additional information directly from hospitals or primary care medical records by the study investigators may be necessary at least for selected major disease outcomes.

### **6.2.3 General Approaches**

As most prospective studies typically include a wide range of health outcomes, it may not be possible to undertake detailed and independent outcome verification and adjudication for each outcome. Particular diseases should be prioritised based on their public health significance, number of cases accumulated, perceived reliability of disease diagnoses, research priorities, and available resources. As a starting point, it may be necessary to focus the work just on several major diseases of global and local health importance (e.g., stroke, IHD, cancer, and COPD). To help inform long-term planning and scope of the work required, it would be prudent to undertake pilot studies, involving a small sample (approximately 500–1000 participants) of randomly selected cases of specific diseases reported across different study areas.

Depending on the findings of these pilot studies, future work may vary, ranging from little extra work (where disease reporting is accurate and information on disease sub-types is adequate), to collection of additional clinical data for a subset of cases reported in particular areas or hospitals, to systematic collection of additional data for most or all such cases over time. For many diseases (e.g., stroke, IHD, and certain cancers), there may be multiple recurrences and readmissions to hospital following a first-ever incident disease event. In selecting relevant cases for verification and adjudication, therefore, due attention should be paid to types of events considered (first-ever incident vs recurrent events). In CKB, several pilot studies were undertaken, involving retrieval and independent review of medical records for ~1000 cases each of stroke, IHD, cancer, diabetes, and COPD (Kurmi et al. 2016). Based on study findings, long-term plans for outcome verification and adjudication have been developed for different diseases in CKB. For diabetes and COPD, the reliability of reported cases was high, and hence, no further verification and adjudication was required. For stroke, IHD, and cancer, which have heterogeneous aetiology, it was necessary to collect additional information from medical records in order to reliably classify major diseases into their aetiologically relevant sub-types.

### **6.2.4 Regulatory Approval**

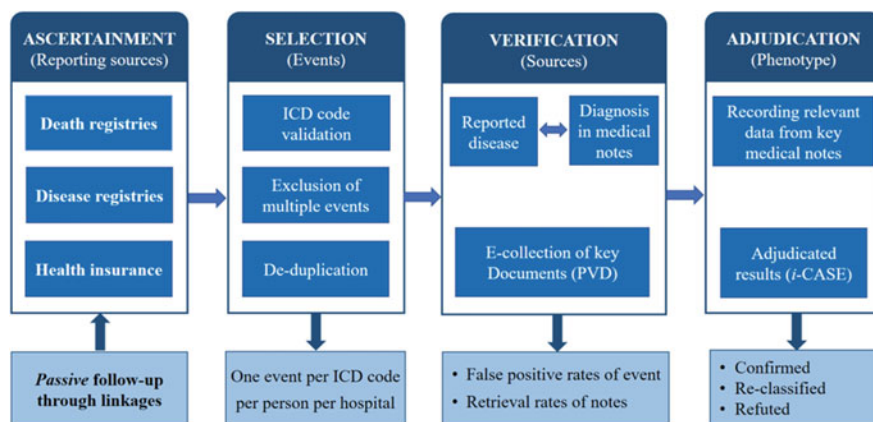
Health records are confidential and access to such data is typically regulated and controlled by national legislation and local regulations (e.g., the UK Data Protection Act 2018). Hence permission to access such records requires appropriate consent of study participants, which is usually obtained at enrolment in the study (see Chap. 2). Moreover, access to health outcomes also requires formal approval by regulatory agencies and governing bodies overseeing such health records, or the healthcare providers who hold such records (e.g., NHS in the UK for the HES and CPRD data). For access to original medical records and electronic health records (EHR), further permission may also be needed from the individual hospitals that store such records.

The process for obtaining approval and access may vary depending on the data sources, sensitivity of the information sought, and local regulations. For cause-specific mortality and cancer incidence that have been widely used by researchers, it may be reasonably straightforward to obtain approval from the relevant agencies in most countries. For other more detailed hospital records, such as HES/HI data and primary care data, it can be more challenging to secure formal approval and access, even in countries with well-established legal frameworks governing their use in medical research. In countries where no appropriate legal frameworks are in place, it may not be possible at all or may take many years of negotiations before formal and limited approval can be granted. To help facilitate the approval process, it may be necessary to limit the data sought only to those that are essential for the planned research (e.g., those related to diagnosis rather than management of the disease and associated costs). Once approved, it is necessary to have a formal agreement with the relevant agencies to specify a range of key issues, including the scope of data to be provided, time schedule, and associated costs.

In CKB, consent to access health record information was included in the written informed consent form approved by the study participants at baseline recruitment between 2004 and 2008. Apart from well-established death registries in study areas, there were also local registries for major disease incidence (stroke, IHD, cancer, and diabetes), with varying degree of completeness and quality. Both registries are managed by local CDC, which are the local study partners, facilitating approval and access. In China, the nationwide health insurance (HI) schemes were launched in 2004–06, which are managed by other government agencies. The CKB obtained permission for linkage with local HI claims databases for study participants, using unique personal identifying (ID) number obtained at initial study enrolment (see Chaps. 3 and 5). Once the linkage is established, it is then possible to access a large amount of information on any episodes of hospitalisations and selected procedures (operations or invasive diagnostic or therapeutic procedures) from HI records. In order to verify and further classify any reported disease diagnoses, it is necessary to retrieve medical records, which is time-consuming and labour intensive and requires permission and cooperation of hospitals. Hence, retrieval of medical records has to be carefully planned and restricted to a limited number of major diseases diagnosed at a relatively small number of major hospitals.

### ***6.2.5 Development of Procedures and IT Systems***

Figure 6.2 illustrates procedures that are typically involved for disease verification and adjudication in large prospective studies. Appropriately designed pilot studies for selected disease outcomes can also facilitate development of Standard Operating Procedures (SOPs) and bespoke IT systems that may be necessary to support and manage large-scale verification and adjudication activities, from identifying and selecting relevant participants, undertaking data collection in hospitals, to secure data transfer and remote outcome adjudication by clinical experts. Table 6.2



**Fig. 6.2** Procedures of outcome verification and adjudication in CKB

**Table 6.2** Examples of IT systems for verification and adjudication of health outcomes in CKB

Stage	Names of software	Functionalities
Planning	Cas Passman NID Check DrList PVD Hospital Standardiser	Manage users for web apps Manage passwords Check validity of national ID number Generate participant list for regional study centre Analyses and match hospital names Automated coding and standardisation of reported outcomes
Verification	LT follow-up PVD Manager Outcome PVD OV-Manager OV-Reports	Collect and manage long-term follow-up events Produce lists of events for verification at local centres Collect source evidence for reported outcomes in hospitals Central management of outcome validation Generate reports of outcome validation
Adjudication	CRD Web <i>i-case</i>	Allocate and manage adjudication tasks to adjudicators Provide internet-based event adjudication

provides examples of bespoke IT systems developed in CKB to support and manage various stages of the work. Apart from software development, the choice of computer hardware device (e.g., laptop, tablet, or mobile phone) needs to be carefully considered. Moreover, the advantages and disadvantages of different IT platforms (e.g., window-based vs web-based) also require careful consideration (see Chap. 7).

### 6.3 Practical Procedures for Verifying Reported Outcomes

In countries such as the UK with well-established nationwide primary health care systems, it is possible to verify reported incident disease outcomes from registries and other health record systems (e.g., HES, cancer registries) using primary health records, which generally contain reasonably detailed medical reports from hospital

for each admission. However, in most low- and middle-income countries without universal primary health care systems, it may only be possible to undertake verification of reported disease outcomes by review of paper or electronic medical records kept in hospitals. This requires corroboration in the hospitals where the patients had been admitted for the particular disease events at the specified date. For large studies involving multiple study areas, this requires feasible and cost-effective approaches to obtain external corroboration of disease outcomes using medical records of the study participants.

### ***6.3.1 Planning and Prioritising***

As prospective studies typically collect a wide range of health outcomes from geographically diverse areas, it would not be feasible nor possible to undertake outcome verification for all types of disease outcomes recorded. In addition to the need to prioritise the diseases, it is also important to plan carefully the number of cases selected and hospitals involved for particular disease outcomes. Initially, it is prudent to focus on 5–10 major diseases of significant population health importance, involving 500–1000 cases for each, preferably selected at random from different areas and across different time periods. If the verification work involves retrieval of medical records, the hospitals should also be carefully selected, taking into consideration the geographical location, types, or ranking of hospitals (e.g., district versus teaching hospitals), permission obtained, and number of cases recorded. For practical considerations, it may be necessary to focus chiefly on hospitals that have recorded a reasonably large number of cases and within easy access by study staff.

The information to be collected from hospital or other sources (e.g., primary health records) for verification will differ greatly between different disease outcomes. Moreover, the information to be collected would also vary greatly depending on whether or not it is combined with outcome adjudication work that would involve independent and detailed review of medical records (see below). In general, specific clinical research forms (CRFs) are required for different diseases, which should be developed after consultation with clinical experts. Depending on work plan for specific diseases, they may involve (1) simple recording of discharge diagnoses without any supporting evidence; (2) collection of discharge summary; and (3) collection of details of presenting symptoms, clinical signs and likely initial and discharge diagnosis, together with discharge summary and specific test reports (e.g., neuro-imaging for stroke, histological tests for cancer types, blood levels of cardiac enzyme and electrocardiograms (ECG) for suspected acute myocardial infarction).



### **6.3.2 *Selecting Diseases***

Since participants typically have multiple disease events recorded during the same or different hospital admissions, and cases for individual diseases may be coded using different coding systems. Before undertaking outcome verification, all disease events ascertained from different sources need to be properly checked and coded using standard coding systems such as the ICD-10 codes. Additional data processing tasks may involve de-duplication (see Chap. 8). It is also likely that many participants may have multiple reports of same conditions over time, which could be new admissions of an old index episode (e.g., multiple admissions for cancer chemotherapy/radiotherapy), episodes of recurrent events (e.g., recurrent stroke), or new onset of same condition (e.g., onset of right breast cancer after left breast cancer). Similarly, many participants may have reports of different conditions at the same calendar date or at some later date. In selecting disease events for verification, it is important to develop appropriate selection criteria for different conditions, distinguishing (1) incident from prevalent cases; (2) fatal from non-fatal disease events; and (3) index cases from comorbid cases. In prospective studies, the verification work should generally focus on the first incident event, or the earliest event for particular conditions recorded as an index case of hospital admission during follow-up. Moreover, as some participants may die at home without proper medical attention, it is more difficult to verify fatal compared with non-fatal disease cases. Some studies may also wish to undertake verification of prevalent cases of certain diseases recorded at baseline for specific purposes (e.g., discovery of genetic variants in genome-wide association studies (GWAS)), which may only be possible through review of primary health care data rather than hospital admission records.

### **6.3.3 *Selecting Hospitals***

Critical to the success of disease validation is retrieval of medical records. This in turn depends upon identifying the individual hospitals where study participants were admitted for particular diseases to be verified. Hospital information can be obtained from multiple sources, including primary care (e.g., family doctor or general practitioner) records, HI records, or death and disease reporting cards. In many circumstances, hospital information may not be recorded properly or consistently, with same hospitals having multiple different names. To facilitate the selection process, it is important to establish a hospital database containing a list of all hospitals involved, including their official names, contact addresses, and official rankings. This can usually be accomplished by searching and downloading official hospital lists from national or local health agencies (or equivalent), updated and supplemented by further checking and modification by the local study staff.

In selecting relevant hospitals, it is important to distinguish hospitals that reported the disease initially (reporting hospital) from one where participants had their

definitive diagnosis ascertained (diagnosis hospital). For the same disease event involving more than one hospital, it may be necessary to select all hospitals involved to increase the likelihood of identifying and retrieving relevant medical records. If the relevant identifiers of the local hospital are missing, it may be prudent to select the highest-ranking hospital. In countries such as China that lack a routine hospital referral system, patients may choose to be admitted to any hospital of their choice. Hence, it may be necessary to identify the hospitals or specialist hospitals where people are most likely to have been admitted if information about diagnosis hospital is unavailable. For example, in CKB, a large proportion of medical records of rural residents could be retrieved from 1 to 2 local county hospitals for most major diseases (e.g., stroke or IHD), while for certain other diseases such as cancer a high proportion of cases were diagnosed at specialised or high-ranking hospitals in nearby big cities. Detailed understanding of the arrangements for local health care delivery can inform planning of outcome verification.

### ***6.3.4 Generating Verification Lists***

To facilitate field work, detailed verification lists should be generated centrally according to specific hospitals involved, containing participants' personal (e.g., name, date of birth, sex, and unique personal ID number) and admission details (e.g., dates of admission and discharge, discharge diagnosis and ICD-10 code, ward name, and admission or hospitalisation number) related to specific episodes of disease outcomes. Prior to the field work, such lists should be sent directly or via regional study offices, using secure methods, to the relevant hospitals so that the medical records can be identified and retrieved in advance. To reduce unnecessary workload, number of site visits, and the need for multiple permissions for access, the work for different diseases should be carefully coordinated and preferably undertaken at the same time. As the verification lists for specific hospitals may contain multiple diseases, the retrieval lists should be prepared to suit to the filing systems of the medical records in specific hospitals (e.g., by disease type, ward number or name, patient name, or admission number).

### ***6.3.5 Organising Fieldwork***

Study staff responsible for undertaking case verification should have basic medical knowledge and be properly trained, especially if it involves retrieval and review of hospital records, to ensure that they comply fully with study procedures and are able to review, select, and record relevant information from the medical records. Before starting any work, it is important to liaise properly with relevant hospital departments to agree on the timelines, work schedules, and methodology for access and extent of such access. To increase efficiency and minimise the time required to

complete the work in hospitals, it is frequently necessary to have two staff members working simultaneously to complete the data collection in a timely manner. Such staff can work collaboratively or independently depending on the specific tasks and procedures involved. Before recording any data from medical records, it is paramount to first check and ensure that the correct medical records for the specific participants requested have been retrieved by the hospitals. For un-retrieved medical records, it is necessary to record possible reasons (e.g., unavailable, filed away permanently, or borrowed by others). For any records that match fully with the study participants, but not with the index events (e.g., different admissions or different diseases), it is still necessary to record details of the retrieved medical records where feasible, but to record information to facilitate further checking and data integration. Depending on the work plan, the data collection may use paper or electronic case report forms (CRFs). Subject to approvals and permission, it is often much more efficient and cost-effective to take photographic images of the relevant pages of the medical records using the built-in camera of computer tablets (or mobile phones) that can be controlled by the bespoke software. In hospitals with proper electronic health records (EHR), it may be more appropriate to simply download the relevant sections of the medical records stored in the EHR.

### ***6.3.6 Processing and Coding Outcome Data***

The information collected should be properly checked, processed, and coded centrally according to the established procedures and coding rules (see Chap. 8). If disease verification involves paper CRFs, then the information collected initially needs to be entered into computer, for which double entry would be standard and can be done in-house or by certain fee-for-service agencies. If verification uses electronic CRFs, then the checking and coding should be conducted automatically at the time of data collection to minimise the need for any subsequent work (see Chap. 2). Although the main objective of case verification is to confirm the accuracy of diagnosis of the reported disease outcomes, information in addition to disease diagnoses collected should also be checked, coded, and integrated into study database to facilitate study monitoring, quality control, data analyses, and different types of research (e.g., better understanding of hospital management for certain diseases).

### ***6.3.7 Case Studies of Verification of Cancer***

In the UK, it is possible to obtain reliable data on cancer types and other relevant information (e.g., histology, key biomarker status) for sub-classification by linkage with cancer registry, HES system as well as primary care (e.g., CPRD) records. Such linkage systems are used to provide confirmation of reported diagnoses of major cancers with very high levels of reporting accuracy. However, such approaches may

not be feasible in other countries without well-established cancer registries. In the CKB, an initial pilot study of ~1000 cancer cases (100 cases from each of ten study areas) reported from local cancer registries and HI systems found that while the accuracy of cancer diagnoses was generally high, little information was typically available on cancer histological sub-types and certain key biomarkers (e.g., oestrogen receptor status for breast cancer, or prostate specific antigen test results for prostate cancers). The results of the pilot study demonstrated the need to undertake a systematic data collection of the histopathological and other test reports through retrieval of medical records. To facilitate the work for cancer (and a few major other diseases: see below), a bespoke IT system, named as “PVD”, was developed for computer tablets, which recorded key clinical data using electronic CRFs, along with photographic imaging of key test reports, relevant for subsequent molecular staging of cancer types (see Fig. 6.3). By the end of 2019, ~20,000 cancer cases had been verified using this approach, showing that for most major cancers (lung, colon, liver, stomach, prostate, oesophagus, and breast), the reporting accuracy of cancer types was high (>90%). Moreover, once cancer histology and other information (e.g., tumour stage, size, and biomarkers) have been collected through PVD, there is little need for additional specialist adjudication of cancer cases. However, for other diseases such as stroke, for which specialist adjudication is needed to reliably classify disease cases into pathologically or aetiologically distinct sub-types (Adams Jr et al. 1993).

## 6.4 Practical Procedures for Outcome Adjudication

Adjudication is the process of independent review of all the available clinical information to confirm and classify reported disease outcomes into relevant disease categories and sub-types according to pre-specified criteria. In randomised controlled trials, it is routinely done to ascertain and improve the accuracy of disease outcomes reported by patients or their doctors. The adjudication requires teams of clinical specialists to review the medical records and reliably confirm and assign cases into major sub-types. As with certain verification studies, adjudication typically requires copies of relevant medical records (presenting symptoms and clinical signs, blood tests, or imaging or other investigations) for the reported cases of specific conditions to further classify the reported diagnoses into sub-types according to pre-specified diagnostic criteria. Depending on the study plan, it could involve a random subset of the cases reported for specific diseases from different hospitals and at different time periods, or all of the cases collected. Although the adjudication work can be done manually, the use of robust and reliable IT systems is necessary to undertake and manage large-scale tasks in an efficient and cost-effective manner.



**Fig. 6.3** Bespoke software to collect information from medical records for cancer in CKB. (a) Generating verification lists for each hospital, (b) Collecting information from medical notes, (c) Capturing data using a built-in camera, (d) Saving data after proper checking

### ***6.4.1 Organising Adjudication Committees***

Adjudication Committees comprise a group of clinical specialists who are appointed to conduct adjudication of diagnoses according to pre-specified criteria. Typically, membership of adjudication committees is composed of independent specialist clinicians that operate independently and are blinded to each other and the additional study data. Membership of such committees should include practising clinicians with expertise in diagnosis and treatment of the relevant diseases and be representative of the types of hospitals used by the study participants. If possible, consultants from high rank hospitals should be invited to be members of adjudication committees. To ensure confidentiality and data security, each member should sign a confidentiality agreement with the study group. The committees can meet periodically in both face to face meetings and electronically to discuss the pre-specified diagnostic and disease classification criteria and work schedules. To facilitate collaboration and effective management of the process, appropriate incentives may be provided to ensure completion of the adjudication in a timely manner. Moreover, members of such committees should be invited to attend periodic local national and international scientific meetings and invited to be co-authors on selected peer-reviewed publications that evaluated their work.

### ***6.4.2 Preparing Disease Cases***

Since the adjudication involves review of medical records, it should be preceded by outcome verification. To ensure consistency and reliability, information collected from medical records related to participant and specific events should be cross-referenced with data acquired from original reporting sources (e.g., HES and HI). Any discrepancies identified should be further reviewed centrally. To minimise unnecessary work by the committees, partially matched records or records with major issues or limited clinical information may not be considered for inclusion in adjudication. Once checked and processed, the retrieved medical records should be integrated with detailed adjudication CRFs for specific diseases to generate adjudication task list for specific members of the committee (see Fig. 6.4). Where possible, any sensitive identifiable personal information (e.g., name, unique personal ID number, hospital admission number) should not be included in the file and the number of cases to be adjudicated by each committee member should be determined according to the agreed schedule and timelines. If the work is done mainly through papers, then the files and detailed medical records should be delivered using special courier service to each member of the committees. If it is done electronically, then the file should be password protected and delivered securely to the committee members. Alternatively, the file can be uploaded into study web servers, which will enable committee members to undertake the adjudication work remotely using different devices without the need to download any documents (see Sect. 6.4.4).



**Fig. 6.4** Bespoke software to undertake clinical adjudication for stroke in CKB. (a) Allocating tasks to different adjudicators, (b) Collecting information from clinical notes, (c) Collecting information from CT report, (d) Confirming and saving results into IT system

### ***6.4.3 Processing, Coding, and Classifying Disease Outcomes***

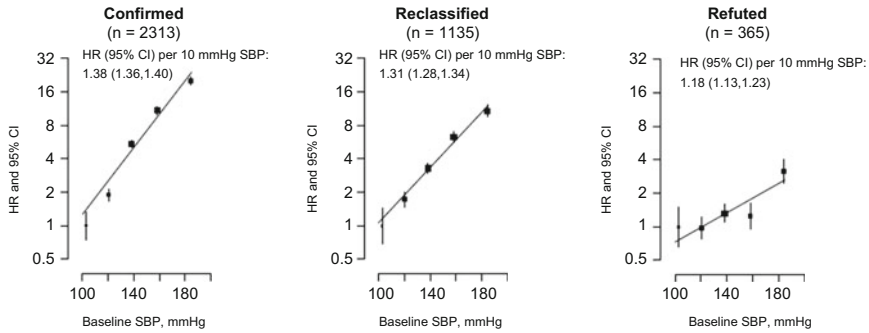
In theory outcome adjudication studies typically have four possible results for each event (i.e., true-positives and true-negatives), for which the assessor correctly judges that the event did or did not occur; and false-positives and false-negatives, for which the assessor incorrectly judges that the event did or did not occur. False-positive and false-negatives will result in bias and systematic underestimation of true associations with diseases. As with disease verification, where it is only possible to collect medical records from disease cases, validation of disease cases yields positive predictive values for disease types and disease sub-types. In large prospective studies, only a small proportion of participants will ultimately develop certain specific diseases, so false positive rates are much more important than false negative rates. Errors in reported diagnoses in cohort studies will lead to bias and underestimation of associations of exposures with diseases and erroneous conclusions about the importance of such diseases. Hence, outcome adjudication should focus on accuracy of case diagnosis rather than on missing cases.

The specific processes required for outcome adjudication may vary for different diseases. For some diseases, it may involve review of medical records and then completing a short CRF with final adjudicated diagnosis according to pre-specified rules and criteria. For other diseases, especially those involving electronic processes, it may also involve further extraction of key data from medical records, followed by final assessment of the disease diagnosis. The latter process, which was adopted in CKB (see below), is more time-consuming, but facilitates quality control, disease sub-phenotyping using certain algorithms (e.g., TOAST criteria for sub-classifying ischaemic stroke), development of automated processes, and new research utilising the data collected (e.g., disease management). To ensure consistency and data quality, the outcome adjudication should be conducted by members of adjudication committees using their native language and checked centrally by research doctors for consistency with international standard criteria and research-specific criteria (see Sect. 6.5).

### ***6.4.4 Case Studies of Outcome Adjudication of Stroke***

In China, stroke is highly prevalent in the adult population for reasons that are still poorly understood. Although the diagnoses of stroke cases mostly involved neuro-imaging even in rural areas, a pilot verification study of about 1000 stroke cases demonstrated that the reported hospital diagnoses of stroke types were unable to classify ischaemic stroke (IS) into LAA or SVD sub-types, or intracerebral haemorrhage (ICH) cases into lobar and non-lobar sub-types. Given the public health importance of the stroke and aetiological heterogeneity of such stroke sub-types, it was decided to adjudicate all the reported first incident stroke cases





**Fig. 6.5** Associations of baseline SBP with ICH risk after outcome verification and adjudication

(along with IHD, cancer, and chronic renal disease) in CKB to enable reliable analyses of prognosis and determinants of IS and ICH sub-types.

In CKB, the stroke adjudication was conducted using bespoke Internet-based Case Adjudication System of clinical Events (*i*-CASE) by Chinese neurologists. This system allowed clinicians to access the relevant medical records allocated to them from their own devices, including desktop computer, mobile phones, i-pads, or tablets, at any time and from any location of their choice (see Fig. 6.4).

Each outcome adjudication typically required about 30 min to complete, involving not only careful review of the medical record images captured by CKB staff through hospital visits but also extraction and entering of selected information from medical records (e.g., those related to presenting characteristics, types of examinations and tests undertaken, use of specific medications, along with reports of neuroimaging or other diagnostic tests or procedures) into adjudication CRF. Based on the careful review of the key medical records, a final adjudicated diagnosis may confirm, reclassify, or refute originally reported stroke types by hospitals, which are further checked centrally as part of quality control process (see below). The detailed information recorded by *i*-CASE also facilitates further classification of stroke types into their main sub-types using established algorithms.

By the end of 2019, a total of 22,700 cases of IS and 3720 cases of ICH had been adjudicated in CKB. Overall, the confirmation rates were 79.6% (95%CI: 79.1, 80.0) for IS and 98.2% (95%CI: 98.1, 98.4) for ICH. For both stroke types, a proportion of the cases were reclassified, mainly from IS to ICH and vice versa, while a small number of cases, often other vascular diseases, were refuted. This has greatly improved relative risk estimates for blood pressure and stroke types (see Fig. 6.5). Moreover, adjudication permitted further sub-classification of IS cases into LAA and SVD IS sub-types using TOAST criteria, and further sub-classification of ICH cases into lobar and non-lobar ICH sub-types.

## 6.5 Monitoring and Data Management

In large prospective studies, verification and adjudication of disease outcomes typically involve many different procedures, requiring detailed organisation, including multiple study and non-study staff. Apart from formal training, development of Standard Operating Procedures and robust IT systems are needed to manage all aspects of the fieldwork, including regular ongoing monitoring of progress, data quality, completeness, and consistency by study regions, by staff and by adjudicators. The key aims of monitoring such data are to detect any issues in the data collection or review process that can be addressed to improve the quality of study procedures. For long-term ongoing adjudication of major diseases conducted over a long period, statistical analyses and periodic reports are required to monitor progress, performance, and data quality using quality indicators, such as retrieval rates or reasons for non-retrieval, by regions or staff members.

Despite the accuracy of diagnosis being the main objective for event verification and adjudication, the consistency of diagnoses is also very important for research, including both inter-adjudicator consistency and the consistency between different adjudicators. Consistency is usually assessed by comparing the adjudicated diagnoses recorded for the same clinical event by two different adjudicators, which can be done centrally as a separate process or built into routine processes involving random subsets of cases (e.g., 10% of the disease events). Moreover, the consistency with international standard criteria also needs to be checked, which can be done by comparing adjudicated diagnoses to a diagnosis generated automatically by computer programmes using built-in algorithms that are developed using standard diagnostic criteria. If the adjudicated diagnosis differs importantly between adjudicators, or deviates significantly from the standard criteria, it should be flagged for further investigation and additional central review by study clinicians.

## 6.6 Summary

In prospective biobank studies improved sensitivity (disease detection) and specificity (disease classification) should enhance the likelihood of identifying important associations with lifestyle, biochemical or genetic risk factors for major diseases. This chapter has highlighted some of the key strategies required for disease outcome verification and adjudication and described practical approaches using examples from work developed in the CKB. Future approaches may involve use of artificial intelligence to standardise ICD-10 codes and develop algorithms for automated disease classification and improved quality control. In particular, artificial intelligence (AI) has the potential to inform disease verification by cross-referencing different health records using key words and to improve detailed characterisation of selected individual diseases using automated algorithms.

## References

- Adams H Jr, Bendixen B, Kappelle L, Biller J, Love B, Gordon D, Marsh E. Classification of subtype of acute ischemic stroke: definitions for use in a multicenter clinical trial: TOAST: trial of org 10172 in acute stroke treatment. *Stroke*. 1993;24(1):35–41. <https://doi.org/10.1161/01.STR.24.1.35>.
- Ay H, Arsava E, Andsberg G, Benner T, Brown R Jr, Chapman S, Cole J, Delavaran H, Dichgans M, Engström G, Giralt-Steinhauer E, Grewal R, Gwinn K, Jern C, Jimenez-Conde J, Jood K, Katsnelson M, Kissela B, Kittner S, Kleindorfer D, Labovitz D, Lanfranconi S, Lee J, Lehm M, Lemmens R, Levi C, Li L, Lindgren A, Markus H, McArdle P, Melander O, Norrving B, Peddareddygarri L, Pedersén A, Pera J, Rannikmäe K, Rexrode K, Rhodes D, Rich S, Roquer J, Rosand J, Rothwell P, Rundek T, Sacco R, Schmidt R, Schürks M, Seiler S, Sharma P, Slowik A, Sudlow C, Thijs V, Woodfield R, Worrall B, Meschia J. Pathogenic ischemic stroke phenotypes in the NINDS–Stroke Genetics Network. *Stroke*. 2014;45(12):3589–96.
- Chen Z, Lee L, Chen J, Collins R, Wu F, Guo Y, Linksted P, Peto R. Cohort profile: the Kadoorie Study of Chronic Disease in China (KSCDC). *Int J Epidemiol*. 2005;34:1243–9.
- 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.
- Green J, Reeves GK, Floud S, Barnes I, Cairns BJ, Gathani T, Pirie K, Sweetland S, Yang T, Beral V. Cohort profile: the million women study. *Int J Epidemiol*. 2019;48:28–29e.
- Herrett E, Shah AD, Boggon R, Denaxas S, Smeeth L, van Staa T, Timmis A, Hemingway H. Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. *BMJ*. 2013;346:f2350.
- Kurmi OP, Vaucher J, Xiao D, Holmes MV, Guo Y, Davis KJ, Wang C, Qin H, Turnbull I, Peng P, Bian Z, Clarke C, Li L, Chen Y, Chen Z. Validity of COPD diagnoses reported through nationwide health insurance systems in the People’s Republic of China. *Int J Chron Obstruct Pulmon Dis*. 2016;11:419–30.
- Lacey B, Lewington S, Clarke R, Kong XL, Xiang L, Chen Y, Yang L, Bennett D, Bragg F, Bian Z, Wang S, Zhang H, Chen J, Walters RG, Collins R, Peto R, Li L, Chen Z. Age-specific association between blood pressure and vascular and non-vascular chronic diseases in 0.5 million adults in China: a prospective cohort study. *Lancet Glob Health*. 2018;6:e641–9.
- Smith M, Zhou M, Whitlock G, Yang G, Offer A, Hui G, Peto R, Huang Z, Chen Z. Esophageal cancer and body mass index: results from a prospective study of 220,000 men in China and a meta-analysis of published studies. *Int J Cancer*. 2008;122(7):1604–10.
- Sun L, Clarke R, Bennett D, Guo Y, Walters RG, Hill M, Parish S, Millwood IY, Bian Z, Chen Y, Yu C, Lv J, Collins R, Chen J, Peto R, Li L, Chen Z. Causal associations of blood lipids with risk of ischemic stroke and intracerebral haemorrhage in Chinese adults. *Nat Med*. 2019;25:569–74.
- Wright FL, Green J, Canoy D, Cairns BJ, Balkwill A, Beral V. Vascular disease in women: comparison of diagnoses in hospital episode statistics and general practice records in England. *BMC Med Res Methodol*. 2012;12:161. <https://doi.org/10.1186/1471-2288-12-161>.

# Chapter 7

## Development and Application of IT Systems in Biobank Studies



Garry Lancaster, Simon Gilbert, and Xiaoming Yang

### Contents

7.1 Introduction .....	147
7.2 Development Strategies and Methodology .....	147
7.3 Development Choices .....	151
7.4 Practical Considerations for Software Development .....	158
7.5 IT Staff and Team .....	162
7.6 Testing .....	165
7.7 Cross-Cutting Issues .....	165
7.8 Summary .....	168
References .....	169

**Abstract** Modern biobank studies tend to be extremely large and complex, and will need to continue for decades. Their success relies critically upon the development and application of comprehensive information technology (IT) systems, not only to collect and store data securely, but also to manage study operations efficiently. Across multiple categories of software, quality and efficiency advantages over traditional paper data collection are potentially profound. However, they can only be fully realised with a carefully planned and organised approach to establishing reliable and coherent infrastructure, selecting or developing software that meets requirements, and choosing the appropriate hardware upon which it will run. Whilst the technical aspects are myriad, the development of biobank IT systems should be centred on people, i.e., those who work on the study and those who participate in it. This guiding principle underpins all decisions throughout the lifecycle, from planning, through development, to implementation. Special care is needed when working with external IT teams, and when integrating specialist hardware into the study. A number of cross-cutting practical issues, such as data and system security, require careful consideration across the whole development effort.

---

G. Lancaster · S. Gilbert · X. Yang (✉)  
Big Data Institute Building, Nuffield Department of Population Health, Old Road Campus,  
University of Oxford, Oxford, UK  
e-mail: [xiaoming.yang@ndph.ox.ac.uk](mailto:xiaoming.yang@ndph.ox.ac.uk)

**Keywords** Biobanks · IT · Software · Hardware · User requirement specification · Outsourcing

## Abbreviations

AI	Artificial intelligence
API	Application programming interface
CAS	Common access system
CKB	China Kadoorie Biobank
COTS	Commodity-off-the-shelf (of software)
CSS	Cascading style sheets
DBA	Database administrator
GDPR	General data protection regulation
HTML	Hypertext markup language
HTTP	Hypertext transfer protocol
HTTPS	Hypertext transfer protocol secure
ICC	International co-ordinating centre
ID	Identifier
IDC	Internet data centre
IT	Information technology
KVM	Kernel-based virtual machine
MMS	Material management system
NCC	National co-ordinating centre
NSIS	Nullsoft scriptable install system
OCR	Optical character recognition
PDF	Portable document format
PHP	PHP: Hypertext processor (a recursive acronym)
PVD	Portable validation device
RC	Regional centre
RFID	Radio-frequency identification
SOP	Standard operating procedure
UI	User interface
URS	User requirements specification
UUID	Universally unique identifier
VM	Virtual machine
XP	Extreme programming

## 7.1 Introduction

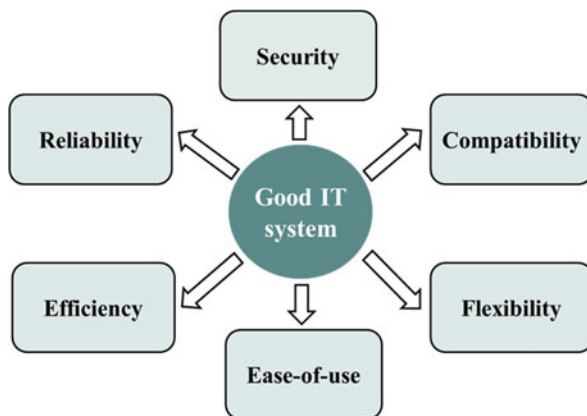
According to the Oxford English Dictionary, information technology (IT) is ‘the study or use of systems (especially computers and telecommunications) for storing, retrieving, and sending information. . .’. These systems are in some cases physical—the computing hardware—yet the majority of IT-related effort is exerted on the non-physical—the programs that run on the hardware, known as the software. The balance of this chapter reflects this, by focussing mostly on software.

Modern biobank studies tend to be extremely large and complex and will last for many years or decades in order to accrue a sufficient number of fatal and non-fatal disease cases to reliably assess their associations with particular risk exposures. Moreover, they may cover multiple geographically dispersed regions or countries, with diverse social, cultural, and linguistic backgrounds. Therefore, the success of such studies relies not only on careful planning and coordination but also on the development and application of robust systems to manage, where possible, all aspects of the study-related activities. Compared to the predominantly paper-driven studies of earlier decades, IT-powered studies provide potentially huge improvements in terms of quality, efficiency, scalability, consistency, security, traceability, and cost-effectiveness. This chapter describes key principles, general approaches, and methodologies for the design and development of biobank IT systems, with practical examples drawn from contemporary large biobank studies. Many of the principles and practical considerations described should also be of general relevance for other population health studies.

## 7.2 Development Strategies and Methodology

IT is a complex and constantly evolving field so the latest information on current technology and future directions should be sought from IT specialists at study planning stage that will not only fit for purposes but also be future proof. Like any large-scale IT effort, development of biobank IT systems involves co-ordinated team efforts of those with different expertise, skills, and roles. IT need not solely passively fulfil requirements considered by researchers; it can also influence study design, organisation, and management, by suggesting new possibilities, and better ways of doing things reliably, efficiently, and cost-effectively. The logical thought processes of IT staff can also contribute more broadly to the biobank project, in key areas such as high-level planning and development of Standard Operating Procedures (SOPs) and requirement-specification documents. Moreover, irrespective of whether they are developed in-house or outsourced, the end product is intended to be used by those who work on the study and those who participate in it, so that, perhaps unexpectedly, many considerations fundamental to such development turn out to be human-related. As well as enhancing capabilities, IT systems typically regulate and control the behaviour of those who use them, which, unless carefully done, has

**Fig. 7.1** Key qualities of a good IT system for biobank studies



the potential to frustrate. Therefore, careful planning and close dialogue and collaboration between IT and other staff throughout the life of the study is crucial, in order to maximise the extent to which each system empowers its operators and helps them to do their work. In general, a good IT system for large biobank (or other) studies should possess a number of key attributes (Fig. 7.1).

### **7.2.1 IT Infrastructure**

Large prospective biobank studies often involve multiple regions or other locations, with complex and differing levels of operational and management structures. At minimum, there will be a top-level (co-ordinating) centre, along with a number of local level centres, although there are many possible variations on this basic pattern. In establishing any IT infrastructure for a study, it is necessary to consider several key factors, including: (1) internet availability, speeds, reliability, and costs; (2) likely quantity of data needing to be collected, transferred, and stored; (3) IT and data regulations in the different locations; (4) physical security characteristics (e.g., door locks, security staffing arrangements); (5) existing local IT infrastructure and staff (e.g., firewalls, IT support staff); (6) requirements for inter-centre linkages, and infrastructure and linkages within each centre. Depending on the study needs and settings, the IT hardware devices required may vary, ranging from mobile phones, through computers (e.g., desktop computers, laptops, and handheld tablets) and servers, to large cloud-based supercomputing and storage facilities (see Sect. 7.3.3).

### 7.2.2 Development Methodology

Software development methodologies vary from strictly ordered stages—the so-called *waterfall method*—to highly iterative *agile* and *extreme programming* (XP) approaches (Beck 2004). XP recommends test-first development, where programmers write test code before they write the code to be tested, and pair programming, where two code together, alternating between ‘driver’ and ‘co-pilot’ roles. Much has been written of the pros and cons of various methodologies, but the key point is to pick one that seems reasonable and then follow it, to avoid chaotic methodology-free development.

In reality, a compromise approach can be productive: the iterative waterfall method (Fig. 7.2), as used successfully by CKB. Like the traditional waterfall method, this breaks down development into distinct stages. However, it differs by

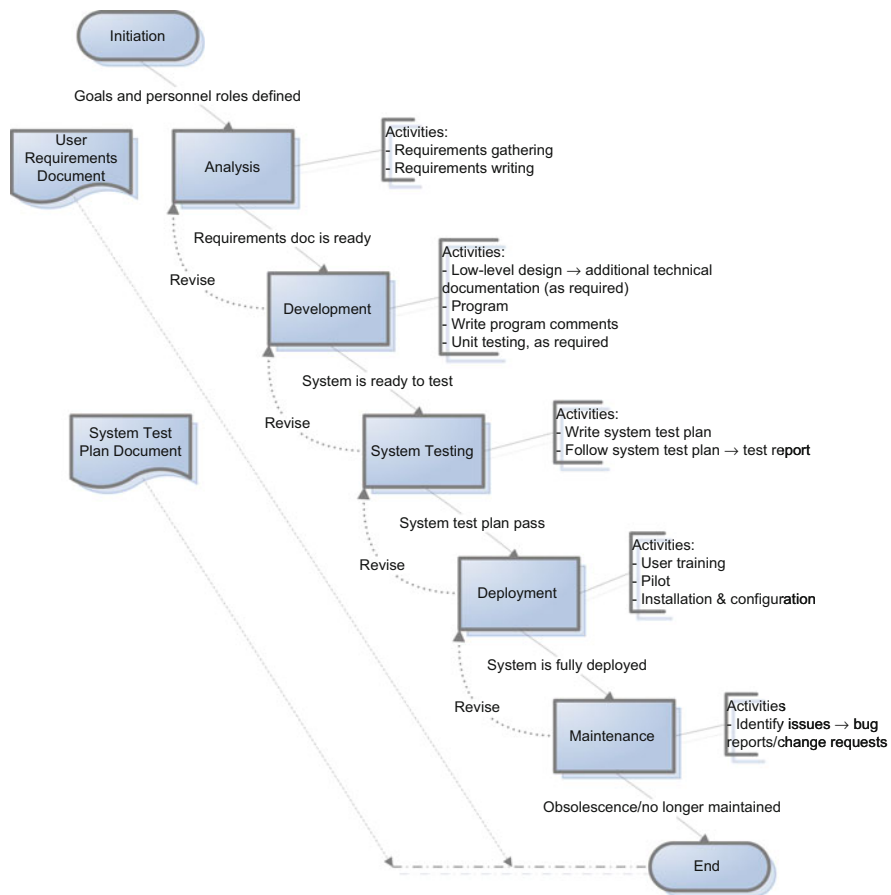


Fig. 7.2 Iterative waterfall process in software development



acknowledging the need for iteration: letting the water flow back up the waterfall. As development flows downwards, the cost associated with any change rockets: as in nature, moving water upwards against gravity is hard work. Part of the skill of software development is knowing when it is time to move to the next stage. To minimise costly future upwards iterations, avoid moving prematurely, but also beware of getting stuck in analysis paralysis.

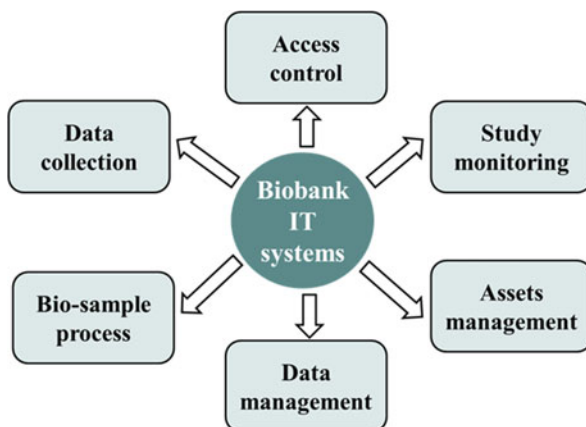
### 7.2.3 User Requirement Specification (URS)

URS is a key document detailing a hybrid requirements and high-level specification for software development. The URS is the *contract*, agreed to by all parties, and is usually written jointly by researchers and IT staff. It is *definitive*, although other (higher or lower level) documents can supplement it, where useful. For in-house development, the study group is responsible for preparing the URS and then following the entire process. For outsourced development, the study group should still actively participate in analysis and production of written requirements, and it is vital to agree an acceptable process with the partner organisation, clarifying respective roles.

### 7.2.4 Categories of Biobank Software

Depending on study design and requirements, a wide variety of IT systems can be developed to support and manage all aspects of study activities (Fig. 7.3). Apart from data collection (e.g., questionnaire interview and physical measurements), they can be used to manage staff at various levels (e.g., access, password), biological samples (e.g., collection, processing, shipment, storage, and retrieval), study materials and

**Fig. 7.3** Main categories of IT system in large biobank studies



devices (e.g., purchase, supply, calibration, maintenance), communication and monitoring (e.g., exchange of message, quality control, statistical monitoring), and systems and infrastructure (e.g., anti-viral, system performance, and maintenance). These systems can be developed under different platform (e.g., web) using different programming languages. Given the range of IT systems that may be required for particular studies, careful planning and prioritisation are essential to ensure they are developed and deployed at the right time of different stages of the project. To start with, systems related to field work may be critical, including data collection, sample processing, staff management, and data transfer, while those related to study monitoring, long-term follow-up, as well as assets management, may be less urgent or critical.

In CKB, over 100 bespoke IT systems have been developed over the years, using different platforms and programming languages. The whole process is carefully planned and prioritised at different stages of the project. Apart from those related to field work data collection (see Chap. 3), sample processing and management (see Chap. 4), and long-term follow-up and outcome verification and adjudication (see Chaps. 5 and 6), a wide range of other systems have been developed to support and manage the study (Table 7.1).


















## 7.3 Development Choices

Once an opportunity for a new project is identified, it is important to develop a detailed plan, through discussions between team members and relevant external parties, to formalise overall development strategies and approaches, and to agree on the key technical requirements in terms of hardware, software platforms, system architectures, and framework. Often a compromise has to be made in order to meet study requirements, timeline, and budget.

### 7.3.1 *Outsourced or In-House Development*














It is likely that, given the highly specialised nature of typical biobank studies, even after a thorough investigation into available Commodity-Off-The-Shelf (COTS) software, as discussed later, there will still be many bespoke IT systems required. Before launching any study, one of the critical IT development decisions, alongside *what* is to be developed, is *who* should develop it. There are two distinct strategies: (1) outsourcing, i.e., pay another (often commercial) organisation or (2) in-house, i.e., directly employ IT development staff. These two approaches have different strengths and limitations (Table 7.2). Depending on the complexity of requirements and timelines, and on the capacities and expertise available within the study team, both approaches can sometimes be combined in the same study (e.g., UK Biobank, which used COTS software to manage biological samples).

**Table 7.1** Examples of IT systems to support study management in CKB

Category	Main functions	
<b>(a) Centre and user management</b>		
	AddRC	List RCs in the study network and add new RCs
	ClinicLoc	Manage clinic areas and locations within an RC
	PassMan	Enable a CKB user to change their password
	UserMan	Allow a CKB administrator to manage CKB users
	CAS	Manage users, groups, authentication, authorisation, etc. for web apps
<b>(b) Monitoring and reporting</b>		
	LatestActLog	View latest IT events including server access, synchronisation, and backup
	LogViewer	View log files sent from study centres to ICC for troubleshooting
	SendLogJob	Send log files of interest to ICC for review
	Reporting	Report recruitment, sample processing, LT follow-up events, etc.
<b>(c) Asset and research management</b>		
	C-FrmTrack	Track the movement of consent forms from RCs to NCC
	Labeliser	Manage the printing of Study ID labels at NCC
	StudyIDAuth	Manage NCC authorisations to print labels for a range of Study IDs
	MMS	Provide service to NCC and RCs for managing consumables and equipment
	CDAS	Provide data access platform for sharing CKB data with researchers
	Standardiser	Analyse, match, and assign ICD-10 codes to diseases in health events
	PVDHospital	Analyse, match, and standardise hospital names in health events
	SpouseChkr	Check the self-reported spouse name against the registry of participants

(continued)

**Table 7.1** (continued)

Category	Main functions	
	Geocoder	Standardise and geocode clinic addresses
	ResearchTrk	Track and monitor CKB research activity including projects and papers
(d) IT system management		
	KeyServer	Provide public key cryptography management service to CKB computers
	LaptopActiv	Activate/deactivate survey laptops within the IT infrastructure of a centre
	Sintegratr2	Authenticate CKB users and start other CKB programs
	UpdateMan	Manage the deployment of software updates to remote study computers
	DbBackup	Backup and restore the local study database of a study computer
	Sinsync	Provide secure asynchronous communication between study computers
	DirSync	Synchronise a directory structure between study computers incrementally
	Sinserver	Provide central file transfer service for Sinsync client computers
	Babel	Localise C++ and Java program with Google translation service backend
	DBTourist	Provide a general-purpose tool for ad-hoc database queries
	NodeStatus	Generate study computers' status reports for remote monitoring

**Table 7.2** Comparison of in-house and outsourced development strategies

	In-house	Outsource
Staff need	Complex	Simple
Initial cost	High	Low
Delivery speed	Slow	Fast
Maintenance	Easy	Difficult
Upgrading	Easy	Difficult
Quality control	Easy	Difficult
Integration	Easy	Difficult
Overall cost	Low	High

If outsourcing, it is important to clarify ownership of the resulting software with the external development partner. Although an obvious goal is to require that ownership rest with the biobank, this is not always necessary or feasible, e.g., with

open-source software development. The main objective is to ensure that the biobank is able to use and maintain any software in the future; whether anyone else also enjoys those same rights is of secondary importance. As well as outsourcing development, it is also possible to outsource ongoing IT support, and the two are often done together using the same partner organisation.

### 7.3.2 Platform Choices (*Web vs Native*)

For large biobank studies involving highly geographically dispersed regions or localities, developing a web-based system may be the preferred option. This involves multiple client computers, each running a web browser, all communicating over the Internet with a single web server, mainly using the HTTPS network protocol. The operators interact with the web browser. The server holds the programs, and runs the server portions directly, delivering web pages (which may themselves contain client program fragments) to the browser upon request. Most web sites use a database for data storage and retrieval, which may either be co-hosted on the web server (Fig. 7.4) or reside on a separate database server. More complex systems may have ‘application layers’ between web server and database layers.

Systems using server components can be designed to utilise cloud computing infrastructure, which provides flexible, abstracted computing services over the

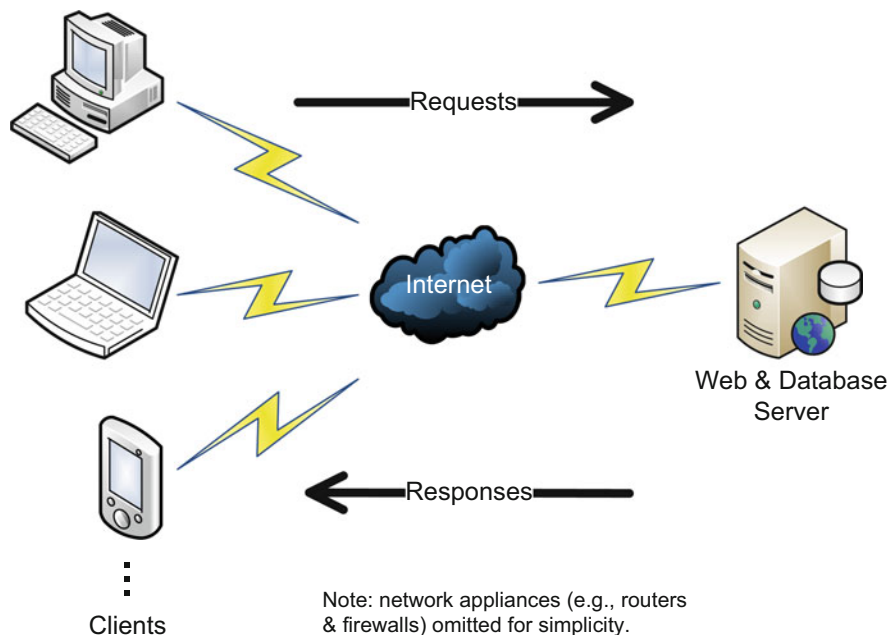


Fig. 7.4 Simple web client/server infrastructure

**Table 7.3** Comparison of web and native development

	Web	Native
Internet connectivity requirements	High	Low
Client compatibility	Wide	Narrow
Need for servers	Yes: at least one web server	Variable: depending on program
Consistency of operator experience	Low	High
Sophistication of operator experience	Low	High
Ease of development	Variable	Variable
Ease of deployment and updating	Easy	Hard
Ease of support	Hard to medium	Easy to medium
Programming languages	Client: predominantly HTML, CSS, & JavaScript Server: popular choices include PHP, Java, Python, & Ruby	Popular choices include C++, Java, C#, & Visual Basic

Internet. Cloud providers take on many of the day-to-day traditional concerns of managing Internet-connected IT infrastructure on behalf of their customers, with NIST listing the defining characteristics of cloud computing as on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service (Mell and Grance 2011).

An alternative approach is *native* development, where programs are installed on the client computers, running outside a web browser (i.e., ‘natively’), giving more flexibility about if, how, and when, they talk to any server computers, and permitting more direct and comprehensive access to the client computer’s resources.

These two approaches have different strengths and limitations (Table 7.3). Depending on the study requirements and settings, both approaches can be used for different activities or operations within the same study. In CKB, for example, certain disease outcome data is collected from hospitals using native programs, but validated by doctors remotely using a web application. Apart from specific requirements, the choice of development platform may also be influenced by existing organisational strengths and expertise, as well as local on-site internet connectivity.

### 7.3.3 Choosing Computer Hardware

Client computers have become increasingly diverse in the 21<sup>st</sup> century. As well as desktop and laptop computers running operating systems such as Microsoft Windows, smaller devices such as tablets and smartphones (most running iOS/iPadOS or Android operating systems) have become increasingly capable of fulfilling useful roles in a study IT infrastructure, as web clients or running native programs.

As for servers, they vary physically from modestly powered devices, resembling desktops in appearance, to hugely powerful rack-mount units. Increasingly, physical servers are being used as hosts for multiple virtual machines (VMs) using virtualisation software (e.g., Oracle Virtualbox, VMware Workstation, KVM (Kernel-based Virtual Machine)). Each VM is largely isolated from other VMs on the same physical host computer; each has a configurable share of processor, memory, disk, and other resources; and each appears to clients as a separate server.

Nowadays even greater storage capacity, abstraction, and dynamic flexibility can be provided by a cloud infrastructure. Whilst it is possible to set up your own, most organisations would turn to existing providers, such as Amazon or Google, offloading the need to purchase, house, and maintain server infrastructure. These service providers give the ability to dynamically allocate resources in response to changes in demand. The disadvantages of using the cloud include the need to work within the cloud provider's framework and a fundamental reliance upon them: any service level guarantees are unlikely to fully compensate for the costs of downtime on a busy biobank project, so this is more about judgements of reputation and trust. Biobanks in particular should ensure they fully understand how and where their data is being stored and transferred by their chosen cloud provider, and confirm that this is compatible with their own legal and regulatory constraints.

### 7.3.4 *Managed Devices vs Bring Your Own (BYO)*

Biobank studies may provide fully managed client computing devices or expect operators to use their own devices. Each approach has its own strength and limitations (Table 7.4). BYO is sometimes clearly indicated (e.g., if requirements dictate that participants will access the study web site or web-based applications from their home computers), while, in other circumstances, it may be obviously inappropriate (e.g., when the requirement is to support only a single model of computer, for use by

**Table 7.4** Comparison of managed devices vs Bring Your Own (BYO)

	Managed device	BYO
Purchase cost	Paid by project	Paid by operator
Ease of device deployment	Variable: hard, if many devices in many locations	Easy
Control over specification	Total	Limited: can set minimum requirements, but not too strictly
Support costs	Low	High
Development costs	Low	High, as need to be more adaptable
Security	High: dictated by project setup	Low: reliant on operators' own setup
Consistency of operator experience	High	Low

staff operators, tightly locked down to exacting security requirements). Clearly, managed devices triumph on all but the first two criteria. However, those two may be decisive.

### 7.3.5 *Software Purchasing Choices*

#### **Box 7.1 Examples of Common COTS Software Relevant for Biobank Studies**

- Operating systems, e.g., Microsoft Windows, Mac OSX, Linux, Android, iOS;
- Office software, e.g., Microsoft Office, LibreOffice;
- Database software, e.g., PostgreSQL (recommended), MySQL, SQLite, Microsoft SQL Server, Oracle;
- Statistical analysis software, e.g., SAS, R;
- Anti-virus and security software, e.g., Windows Defender, Kaspersky, Sophos;
- Software development tools, e.g., Embarcadero C++ Builder, Microsoft Visual Studio, NetBeans;
- Version control software, e.g., Git, Subversion, Mercurial;
- Low-level documentation generation software, e.g., Doxygen, JavaDoc;
- Bug tracking software, e.g., Bugzilla;
- Installation software, for building installers e.g., NSIS, InstallShield;
- Translation software e.g., Multilizer, Sisulizer, Qt Linguist.

It is usually far cheaper and more convenient to use something that already exists than to create it specifically for a project. It will always be possible and necessary to use some COTS software (see Box 7.1). Although much COTS software is commercial, some is available free of charge, as open-source or freeware. Whilst it can be tempting to assume that *free* software must be *worthless* software, in fact it can be of extremely high quality. Open-source software licences often enforce obligations to distribute the software source code (i.e., programs in their original textual form, as written by developers) upon request; this can even apply ‘virally’ to other custom software that integrates with the open-source software. In any case, studies should carefully check the licence agreements of *all* software before use.

Some commercial COTS software is available with special low academic pricing, available to many organisations setting up biobanks. If none is obvious, it can be worth contacting the software producer directly, as companies are often happy to help socially beneficial projects, perhaps in return for some kind of acknowledgement of their assistance. Even if it appears that no COTS software satisfies requirements, it is still prudent to check whether there is any that *almost* does, in which case,



perhaps through small adjustments to non-critical requirements, it may nonetheless be a valid option.

## 7.4 Practical Considerations for Software Development

There are certain principles and practical considerations that should guide IT development work, from project initiation onwards. Whilst some may appear to be common sense, they are often overlooked and inconsistently applied.

### 7.4.1 *User Friendliness*

Whilst the technical aspects are myriad, the development of biobank IT systems should always be centred on people, i.e., those who use them. Large biobank studies tend to involve many *operators* who will operate the IT systems, and their interests and levels of IT knowledge and competence must be carefully considered. Depending on the study approach, the operators may be well-trained study staff yet, in some cases, they may be study participants themselves, e.g., as with the self-administered questionnaire and cognitive tests of UK Biobank (UK Biobank 2007). Systems designed for these ‘participant operators’ need to be extremely easy to use for those with varying levels of computing knowledge, without the benefits of prior training. The remaining study personnel, in the IT team and more widely, also have important requirements that need to be satisfied to ensure success.

### 7.4.2 *Simplify*

In a project of any complexity, simplicity pays off in the long-term. To achieve this, it is worth spending time early on contemplating designs, deciding which features are essential, which are highly desirable, and which are neither. The ultimate achievement in this respect is to identify something as redundant and remove it entirely; it is clear that code that does not exist can *never* have problems. As well as simplifying code, the process involved should also be simplified to aid visibility and optimise working efficiency.

There are two ways of constructing a software design: One way is to make it so simple that there are obviously no deficiencies, and the other way is to make it so complicated that there are no obvious deficiencies. The first method is far more difficult. C. A. R. Hoare (Hoare 1981)

### 7.4.3 *Be Conservative*

Try to stick to approaches already proven in the field because, although novelty is intellectually stimulating, it is high risk as well. However, when nothing exists that fits, this is permission to try something new. So, a general approach is to try and test where you can, and go for novelty where you must. For example, CKB waited several years before adopting new versions of Microsoft Windows, preferring older, well-understood versions over inhabiting the cutting edge.

### 7.4.4 *Carefully Preserve Data*

Biobank studies should maintain a full data history, to satisfy external auditors and to provide a way of tracing all data back to the ‘source data’—the original data, as it was first collected or imported into the study, even if in a format unsuitable for analysis (see Chap. 8). To enable this, studies should not routinely delete or modify data. Instead, new versions of data ought to be created to record each successive change. Whilst the latest version will be the most used, and definitive, the complete history is preserved for when it is needed (see Box 7.2 for a technical overview). One exception applies if participants request removal of their data from the study, in which case such requests must be honoured with real data deletion, while keeping a clear record of specific actions taken.

Alongside the core data, details of *who* made the change and *when* it was made should also be recorded by the system. Where this is not implicit, further audit data such as *where* the change was made, *what* system made it, and *why* it was made should also be recorded.

#### **Box 7.2 Database Design Techniques for Data Preservation**

- Generally avoid use of database DELETE and UPDATE statements: use INSERT instead, plus a version ordering field, such as a timestamp;
- Indicate logical record deletion by setting a special ‘virtual deletion’ flag field;
- Use rules and views to ensure that such data is straightforward to work with, e.g., so that the set of most recent rows for each entity can be simply obtained.

Data must be regularly backed up. It is necessary to perform test restores of backups to ensure that backup procedures are sound and reliable. Some backup files can be stored nearby for quick retrieval, while others can be stored off-site (in case of site-wide disaster). If using database replication (that is, multiple databases that synchronise data with each other), this should not be used as the sole means of backup, although it offers additional redundancy.

### 7.4.5 Monitor All Systems

All IT systems for the study should be diligently monitored, especially those that affect safety, are responsible for critical data changes, are expected to be encounter of frequent external problems, or are required for regulatory compliance. One valuable monitoring technique is to have all systems write log files as they run. Logs should contain anything that might help IT support staff diagnose problems (e.g., progress through the program, details of any errors), with the exception of passwords, personal data, or anything else potentially sensitive. In CKB, a program SendLogJob runs periodically on each computer across multiple study locations to check every log file, sending those that are ‘interesting’ (i.e., containing certain trigger words like ‘ERROR’) to the IT team for human inspection, aided by the LogViewer program (see Fig. 7.5).

Projects should have a set of reports, available to the responsible staff across the project, with some utilising a simple, ‘at a glance’/dashboard-style, for the most crucial metrics, highlighting problems, perhaps using traffic light style indicators, to prompt further review and quick action.

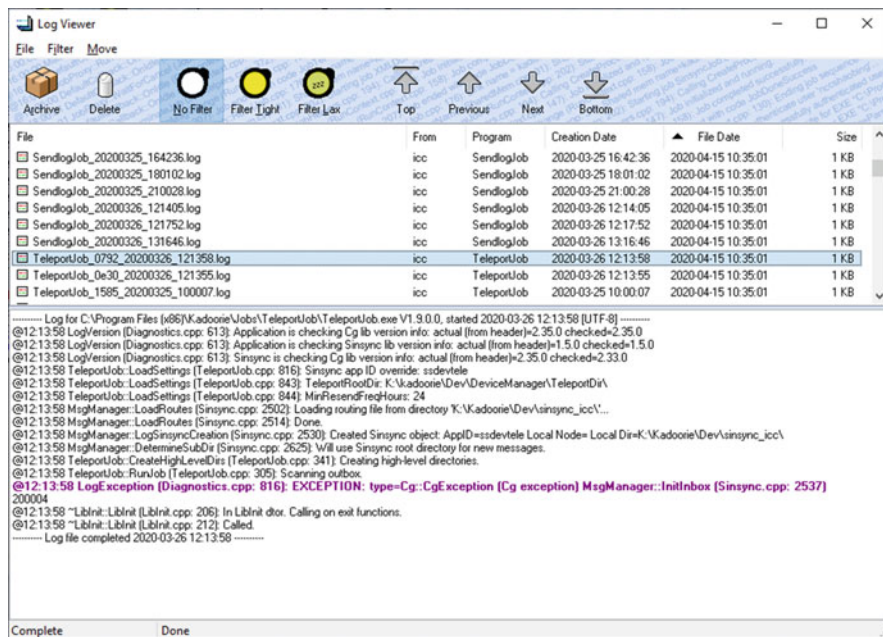


Fig. 7.5 Example of LogViewer screenshot in CKB

### 7.4.6 *Maintaining Security Across Different Systems*

While designing and developing IT system for biobank studies, it is paramount that data and system security can be properly maintained throughout the life of the study and across multiple study locations and operators (see Sect. 7.7.1 for further details). The proper provision of data and system security will: (1) help maintain privacy of personal and confidential data; (2) protect the integrity and reliability of data, by preventing modifications both from malicious actors and from accidents; and (3) adhere to relevant governance rules and data protection laws, e.g., the General Data Protection Regulation (GDPR).

### 7.4.7 *Reuse Appropriately*

The creation and repeated use of reusable software components, for example, in the form of software libraries, or small ‘building-block’ programs, avoids duplication of effort, and aids adherence to shared conventions and standard. However, there are additional costs associated with reuse of code/libraries (approximately three times the programming effort of single-use code). As highlighted by Kevlin Henney, ‘...there is no such thing as reusable software, only software that has been reused’ (Henney 2002). If reuse is clearly indicated, that is one thing, but speculatively designing for reuse is wasteful, as often the speculation is wrong. In the worst cases, code designed for reuse is never used *at all*—code must first be used before it can be reused.

After a brief initial assessment, proposed units of code (e.g., classes or modules) can each be assigned to one of three categories of reuse. This classification should determine the development approach as follows:

1. Definitely no reuse: No need to put in a library or building-block program. Develop, document, and test solely in the current context of use;
2. Definite reuse: Place in a library or building-block program. Develop, document, and test for a comprehensive range of valid usage;
3. Not reused now, but may be in future: Proceed as point (1), except with one extra consideration—whenever making an implementation choice that has little cost penalty either way, select the option that is most amenable to future adaptation for reuse.

### 7.4.8 *Prototyping*

If analysis of development requirements gets bogged down, it can be helpful to consider prototyping: writing a quick program to demonstrate part of the functionality (e.g., a UI prototype without model or database code), for people to explore,

review, and refine. Practical demonstration often provokes insight: sometimes you only know what you want when you do not get it. Prototypes can be either evolutionary (intended to be enhanced later to full functionality) or throwaway (intended to be used as a prototype then discarded), and it is crucial to be clear about which kind is being developed before starting, as different quality standards will apply throughout development. Throwaway prototyping may appear wasteful, but if there is great uncertainty it may be very efficient, as it is quick.

### **7.4.9 Translation**

In studies involving multiple countries or populations with different linguistic traditions, it is often necessary for software to support multiple human languages, or a single language which is not spoken by the developers, in which case translation issues need to be considered. User interface (UI) translation is known as *localisation* and, to facilitate the localisation process, developers must adopt certain conventions when coding it, a discipline known as *internationalisation*. The internationalised software is passed on to translators, who then localise it using a suitable tool, e.g., Multilizer. Translation of some IT-related documentation, such as online help, may also be required.

## **7.5 IT Staff and Team**

IT development is a complex and highly specialised process, requiring people with different backgrounds, training, skills, and experiences.

### **7.5.1 Roles of Different IT Staff**

IT staff are heterogeneous, with many specialist roles (Table 7.5). To ensure that research objectives prevail, all IT staff are ultimately under the direction of non-IT project leaders and should work closely with other non-IT staff as part of the same research team.

In practice, many IT staff combine specialisations, it being especially common for notional developers to have a wide remit. One combination that ought to be avoided, if possible, is that of developer and tester, at least for any single program. It is essential for each to maintain their own understanding of what each program ought to do (even though both will have read the same specification documents), so that the misapprehensions of either party can be more readily identified and then corrected.

**Table 7.5** Common names of IT role and specialisations

Role name	Role description
Project manager	Manages other IT staff and day-to-day coordination of IT projects
Analyst	Liaises with non-IT staff familiar with the problems being solved, to tease out and document requirements, and develop high-level specification
Software architect	Develops user requirements and lower-level specification documents for software design, guiding developers on implementation
Developer	Writes and maintains software that satisfies specification
User interface (UI) designer	Designs the user interface ('look and feel') of software, producing UI design documents, or easy-to-integrate UI language 'design code'
Tester	Produces test plans, tests software against both specification and commonsense principles, and writes test reports detailing results
Technical writer	Writes online help and other software documentation intended for operators
Database architect	Designs database tables and other database schema elements
Database administrator (DBA)	Installs, configures, and maintains database systems, including managing backups
Data scientist	Curates, transforms, and integrates collected data, making it available in a form suitable for research
Systems administrator	Helps select hardware, and commodity-off-the-shelf (COTS) software, and installs, configures, and maintains it, including backups
Helpdesk operative	Provides first line support, resolving basic issues, and passes more involved cases to appropriate team members
Translator	Translates software user interfaces (UIs), and some documentation, from one human language to another, e.g., from English to simplified Chinese

### 7.5.2 *IT Team Communication*

Much of IT work is novel and non-obvious, which differentiates it from typical factory work (see Box 7.3). Unfortunately, this comes at a cost: if each person has to communicate with every other, the number of communication paths increases exponentially as team size increases.

#### **Box 7.3 The Factory**

In a factory in which physical items are built, if one factory worker builds ten items in an hour, then two workers could make the same in half an hour. Production line work is easily scalable; what needs to be done is well-known, as countless identical products have been made before; constructing yet another requires neither innovation nor decision making; crucially, it does not require the different workers to *communicate*.

As such, there are major challenges in undertaking and managing a major IT project (Brooks Jr. 1995). In one memorable observation, Brooks states that, ‘adding manpower to a late software project makes it later’, because additional communication overheads swamp any realistic benefit. He recognised the significance of streamlining communication and distributing relevant information efficiently. Much of this is through creating written documentation, providing a permanent record of important decisions, avoiding the need for repeating the same exchanges and promoting evolution of ideas.

Since the publication of Brooks’ work, technology has advanced substantially. There are now many tools to aid communication throughout a project, for example:

- Version control systems streamline developer cooperation in creating the source code for a single system;
- Sophisticated word processors speed the creation of specifications and other technical documentation;
- Diagramming tools allow the creation of powerfully informative flowcharts and more specialist software diagrams, e.g., database schema designs;
- Bug tracking tools allow the management of software defects and change requests.

By themselves, these do not solve the problems of communication—no mere tools can do that—but they can be used intelligently as part of the solution.

### **7.5.3 Working with Other IT Teams**

Working with other IT teams outside the biobank project team, either within the same organisation or externally, presents challenges. To smooth interactions, establish clear responsibilities and communication paths at management and technical levels at an early stage.

It is crucial to agree a simple, well-documented, way of transferring data between the two teams’ software (an *interface*), e.g., library API, file format, database schema, network communication protocol. Clearly delineating the dependencies between the two teams allows each to work independently on their ‘end’. Testing the two sides’ work together, early and often, will pay dividends in sorting out any incompatibilities before they become ingrained.

The same approach can be used internally within larger biobank IT teams, to sub-divide work and rationalise communication paths.

## 7.6 Testing

For quality assurance, every system must be properly and thoroughly tested before rolling it out for formal use. Although developers should test their programs as they write them, and judiciously create test code to exercise features ('unit tests'), there should be a separate tester for system testing. Testing should aim to identify bugs (i.e., software defects) of two kinds: (1) the program does not follow its specification and (2) the program does not follow 'common sense', regardless of what its specification says. The latter kind can sometimes cause disagreement, but it is inevitable that not everything is captured in a specification. Most bugs are resolved by modifying the program, but other resolutions are sometimes appropriate, e.g., specification changes. For each bug, a useful bug report should be filed, including:

1. A clear description of all steps required to show the problematic behaviour;
2. What happened that was unexpected;
3. What should have happened instead;
4. References to any parts of the specification that were violated, where appropriate.

Bug reports, and their follow-up, need to be managed carefully with a program like Bugzilla, to ensure all the issues raised will be reviewed and dealt with properly by the relevant people. Information about the underlying causes or possible resolutions may be valuable, but this is not the primary concern of testers. Everyone involved in testing should refrain from emotive, adversarial language, or assigning personal blame and, instead, concentrate on their shared endeavour of using the process to maximise quality.

## 7.7 Cross-Cutting Issues

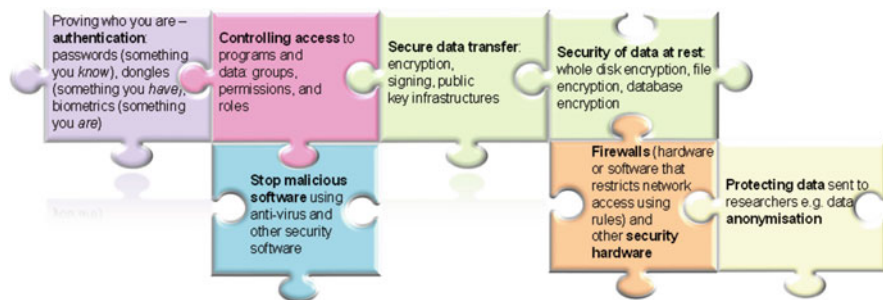
There are a number of cross-cutting issues that ought to be considered throughout the development of biobank IT systems, starting from the abstract analysis stage but, especially, onwards throughout implementation and beyond, the most critical of which are detailed here.

### 7.7.1 Security

IT team members should familiarise themselves with the technical aspects of security, and have a basic understanding of legal and regulatory aspects. They should communicate these matters effectively throughout the team, when appropriate.

Especially for systems exposed externally (e.g., over the Internet), it is worthwhile undertaking threat analysis: from whom (or what) does the data need to be protected? As well as considering malicious actors (external or internal), it is also





**Fig. 7.6** Security jigsaw concerning IT systems

useful to consider how accidents could compromise security, due to either human error or random events. Aspects of security can be depicted as a jigsaw (Fig. 7.6).

Over the years, experts have developed specialised and effective cryptographic algorithms (e.g., for encryption and signing). There is, therefore, no need at all to ‘roll your own’, which will cost resources and compromise system and data security. Where possible, COTS security software of good reputation (e.g., VeraCrypt, Crypto++) should be used (‘full stack’ where possible, otherwise libraries of building blocks), falling back on respected print sources (Ferguson and Schneier 2003; Press et al. 1992) if necessary. As errors can creep in inadvertently, implementations of standard algorithms should be tested with published input/output pairs to check correctness.

The decommissioning of any study IT devices should also be handled carefully. Prior to equipment disposal, 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 for future auditing and backup purposes. All confidential information should then be securely erased and physically destroyed, with a record of the destruction of devices and data properly logged for future reference.

### 7.7.2 *Choice of IDs and Codes*

Everything in a project needs an identifier (ID) or code: a name to be known by. Choosing effective IDs is crucial. The ID of each object must be strictly unique, at least within all other objects of the same type, but possibly across all objects in the project or even globally, as with a UUID (Leach et al. (2005)). Although they may have other uses, these unique IDs are used within database systems as primary keys, which each identify a single row in a database table. IDs ought also to be immutable, i.e., once an object has one, it should have no other.

**Box 7.4 Issues with English Characters in IDs**

- Across the world, people's level of familiarity with English characters varies.
- Case issues: will you use lower case, upper case or both? Databases and other software vary with respect to case-sensitivity (whether different cases of the same character are considered different or the same) and sort orders.

Identifiers may be either pre-existing ('natural') or generated and assigned by the IT system ('artificial'). Use of 'almost unique' natural keys, like people's names, or most countries' social security numbers, should be avoided because these are sometimes reused, duplicated, or changed, making them unsuitable as IDs. Often, IDs will be purely numeric, but many also use alphabetic English characters, which can create additional issues (Box 7.4). Going further, so that *non-English* letters are used in IDs, introduces further human and IT issues, and should only be contemplated if they are otherwise-excellent natural IDs.

IDs are frequently present in the physical world, so need to be entered by operators into IT systems. Short and easy-to-enter IDs will save time and reduce error rates (although it is wise to make them long enough to leave room for expansion over initial requirements, e.g., by adding one or more extra characters). To minimise human effort and error, barcodes or RFID (radio-frequency identification) tags encoding the IDs can be attached to the corresponding objects, if appropriate, so that the IDs can be entered automatically by a scanner or reader. Any ID that does need to be typed (even if only in a fall-back scenario, when a scan or read fails) ought to have a final check digit or character to detect errors, e.g., a Verhoeff check digit (Press et al. 1992) as used by CKB (see Chap. 2) (although developers should be aware that there are at least two incompatible variants of this algorithm in widespread use). Where objects of different types occur together, confusion can be minimised by employing distinct ID schemes for each type, e.g., with different lengths, formats, or prefixes.

### 7.7.3 Computer Clocks

Accurate computer clocks turn out to be surprisingly important for security and consistency, e.g., to ensure accuracy of audit data and to record time ordering of data correctly. Most computers keep good time when powered on, and when they are switched off there is a button cell battery in each desktop computer that keeps the clock ticking. Although the battery should usually last for years, unfortunately it sometimes fails early, so any study computers with an Internet connection should be configured to automatically set their clock from a public time server. For offline computers, it is necessary to enforce regular manual checking of date and time by operators, e.g., by running a small program at each start-up that prompts them to

check. Even well-regulated clocks may drift a little, so study software and operations must accept minor discrepancies. In CKB, for example, clinic laptops rapidly exchange time-stamped ‘heartbeat’ messages, and they accept timestamps between 10 min old and 5 min into the future.

#### **7.7.4 Working with Specialist Hardware Devices**

Most biobank studies tend to collect large amounts of data using specialist hardware devices, such as blood pressure monitors, lung function devices, and bio-sample analysers. Comprehensive integration of these devices into the biobank IT systems will minimise data transcription errors and simplify the operator and participant experience.

Before committing to any particular device, it is advisable to seek comprehensive information from the manufacturer about its technical aspects. Most devices are not intended for use in research settings, so careful evaluation of documented hardware or software interfaces (e.g., APIs, serial port communication protocols) will be needed to inform decision making (including purchasing decisions) and to form the basis for integration efforts.

If potentially utilising a manufacturer-provided UI, it is also necessary to check its suitability including, for multilingual biobanks such as CKB, its support for multiple languages. In some cases, it may be possible to modify the manufacturer’s software to add additional language support, provided they grant permission. As many studies continue for years or decades, it is also vital to consider issues related to ongoing support, e.g., cost, duration, and geographical coverage.

### **7.8 Summary**

This chapter provides a high-level overview of the main strategies, methodologies, and practical considerations concerning development of IT systems and infrastructure for large biobank studies. Our experience is based mainly on the CKB study, for which well over 100 bespoke IT systems were developed in-house by the project IT team, covering all aspects of the project. The successful development and application of these systems have transformed the way the study is being run and managed. Moreover, they have greatly improved the quality and completeness of data collected, management efficiency, and cost-effectiveness. Although these IT systems are unique in many ways, they do provide an instructive example and a starting point from which future studies may learn and benefit, irrespective of their design, size, complexity, and settings.

## References

- Beck K. *Extreme programming explained: Embrace change*. 2nd ed. Upper Saddle River: Pearson Education; 2004.
- Brooks FP Jr. *The mythical man-month: Essays on software engineering*. Anniversary ed. Boston: Addison-Wesley Longman; 1995.
- Ferguson N, Schneier B. *Practical cryptography*. Indianapolis: Wiley; 2003.
- Henney K. (2002). The imperial clothing crisis. Retrieved from <http://www.two-sdg.demon.co.uk/curbralan/papers/minimalism/TheImperialClothingCrisis.html>
- Hoare CAR. The Emperor's old clothes. *Commun ACM*. 1981;24(2):75–83.
- Leach P, Mealling M, Salz R. (2005). A Universally Unique Identifier (UUID) URN Namespace, RFC 4122. <https://doi.org/10.17487/RFC4122>
- Mell P, Grance T. *The NIST definition of cloud computing*. (800-145). Gaithersburg: National Institute of Standards and Technology; 2011.
- Press WH, Flannery BP, Teukolsky SA, Vetterling WT. *Numerical Recipes in C*. 2nd ed. Cambridge: Cambridge University Press; 1992.
- UK Biobank Limited. (2007). UK Biobank: Protocol for a large-scale prospective epidemiological resource. Retrieved from <https://www.ukbiobank.ac.uk/wp-content/uploads/2011/11/UK-Biobank-Protocol.pdf>

# Chapter 8

## Management and Curation of Multi-Dimensional Data in Biobank Studies



Gary Sansome and Alex Hacker

### Contents

8.1	Introduction .....	172
8.2	Guiding Principles for Data Management .....	173
8.3	Frameworks for Data Management .....	174
8.4	Data Gathering and Validation .....	178
8.5	Data Cleaning and Standardisation .....	183
8.6	Data Linkage and Integration .....	185
8.7	Data Aggregation .....	189
8.8	Quality Control and Documentation .....	192
8.9	Data Integration Platforms .....	194
8.10	Governance and Access to Data .....	197
8.11	Summary .....	201
	References .....	202

**Abstract** The development of secure and reliable systems to collect, store, utilise, and share data on study participants plays a critical role in large population health studies. Contemporary prospective biobank studies typically involve hundreds of thousands of participants, and collect a wide range of data through questionnaires, physical measurements, sample assays, and linkages with external data sources for an extended period. Careful planning and management of a central data repository are required to ensure the privacy, security, accessibility, flexibility, consistency, and accuracy of the data collected and generated in the study. This chapter outlines some of the key concepts and principles underlying the design and development of data storage infrastructures, database architecture, and management systems in large biobank studies. It also describes practical considerations for each step from initial data collection from study participants to delivery of research-ready datasets; from

---

G. Sansome (✉) · A. Hacker

Big Data Institute Building, Nuffield Department of Population Health, Old Road Campus,  
University of Oxford, Oxford, UK

e-mail: [sam.sansome@ndph.ox.ac.uk](mailto:sam.sansome@ndph.ox.ac.uk)

© Springer Nature Singapore Pte Ltd. 2020

Z. Chen (ed.), *Population Biobank Studies: A Practical Guide*,

[https://doi.org/10.1007/978-981-15-7666-9\\_8](https://doi.org/10.1007/978-981-15-7666-9_8)

171

data import, cleaning, and integration; through quality checks, standardisation, and validation; and finally to preparing datasets for *bone fide* researchers. The general principles and approaches described should be applicable to a wide variety of population health studies in different settings.

**Keywords** Biobank · Big data · Standard operating procedures · Data management · Data sharing

## Abbreviations

API	Application programming interface
CKB	China Kadoorie Biobank
DAG	Data access governance
DBMS	Database management system
ICD	International classification of diseases
ID	Identifier
IT	Information technology
RDBMS	Relational database management system
SQL	Structured query language
SOP	Standard operating procedures
WHO	World Health Organisation

## 8.1 Introduction

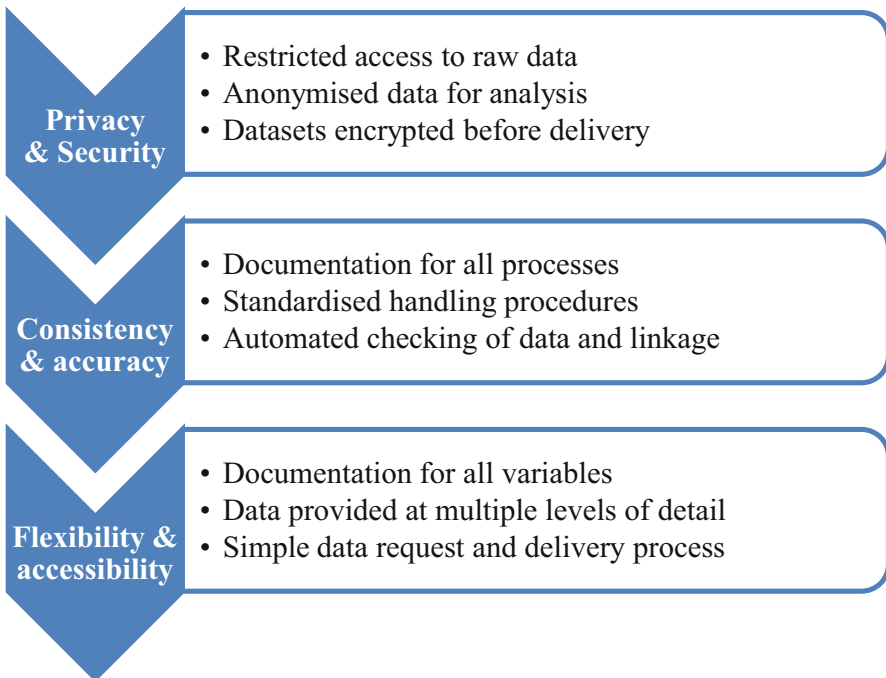
Historically, prospective studies collected a modest amount of data from a small number of participants, with limited follow-up, and the data generated were generally used by relatively few researchers within a particular organisation. As such, simple procedures and tools sufficed to collect, process, and manage the study data effectively. By contrast, contemporary prospective biobank studies typically recruit many more participants and gather a wide variety of data from diverse sources, including not only exposures (e.g., questionnaires, physical measurements, and blood assays), but also data from external sources (e.g., hospital records, primary care, and air pollution monitoring records). Such expansion requires secure, robust, and efficient data management systems to collect, process, store, manage, and use the data, not only internally but also externally with collaborators and other verified researchers. Moreover, rapid advances in biotechnology have greatly enhanced our ability to turn samples into data (e.g., whole genome sequence), generating and linking unprecedentedly large and complex datasets. These require novel approaches, infrastructures, and systems to process, manage, and use these data. This chapter describes key principles and practical considerations for the design, development, and implementation of data management systems, with practical examples drawn from contemporary large biobank studies. Many of the principles

and procedures described should also be applicable to other population health studies (e.g., cross-sectional surveys, case-control studies).

## 8.2 Guiding Principles for Data Management

The key requirements for data management in large prospective biobank studies are to ensure data privacy, security, accessibility, flexibility, consistency, and accuracy of any data collected or generated (Fig. 8.1). To fulfil these criteria, robust data management systems and procedures should be developed that cover all aspects of the study data, from collection to linkage, integration, storage, and use. Increasingly, biobank studies require reliable platforms and procedures to share their data with the wider scientific community. Depending on the size and complexity of the data, the IT platform required for storage, processing, and analysis may be a conventional desktop computer, an internal client-server configuration, or a cloud-based supercomputing and storage service.

The paramount requirement for data management is to protect data *privacy* and ensure the *confidentiality* of study participants. Any personally identifiable information (e.g., name, address, telephone number, and IDs) should be stored *securely* and



**Fig. 8.1** Core data management requirements in biobank studies

separately from other data and be accessible to only a very limited number of key study staff. Any data made available for analysis should be anonymised and any data transfer outside the network's firewall should be encrypted. Given the complexity of the tasks involved in data collection and management for large biobank studies, systems and procedures should be in place to ensure *consistency* and *accuracy* of the data collected and managed. Collecting and processing data in a standardised and well-documented way will minimise the risk of errors. Despite these efforts, mistakes and inconsistencies are inevitable, so appropriate procedures must be developed for identifying and addressing issues at every stage of collection, linkage, and integration. In prospective studies, false positives are more serious than false negatives, especially those related to health outcome data. For example, given a cancer diagnosis with unclear patient details, it is more appropriate to ignore it completely than to assign it to the incorrect participant. While missing data reduce the power of a study, incorrect data risk yielding erroneous results (see Chap. 1). Finally, the more *flexible* and *accessible* a dataset is, the more valuable it is. This requires well-chosen and well-documented data, and a technical and logistical infrastructure designed to deliver datasets rapidly and securely.

### 8.3 Frameworks for Data Management

Large biobank studies usually involve many different types of data, collected and generated at different stages, by different means, individuals, and organisations, with varying degrees of complexity, completeness, and quality (Fig. 8.2). Combining these into a research-ready database requires a wide selection of resources, including: (1) documentation, from the high-level data management plan to detailed standard operating procedures (SOPs); (2) suitable hardware, software, and systems architecture to store, process, and deliver the data; and (3) of course, staff.

#### 8.3.1 Data Management Plan

A data management plan is a key document that describes the types of data that are to be collected, transferred, checked, integrated, and stored, and how the results are to be accessed. Components of a data management plan should include: (1) details of manual and automatic validation; (2) arrangements for handling data irregularities; (3) procedures for modifying data after collection; (4) rules for updating personal information; and (5) specification of the location for storage and integration of raw data files. A data management plan needs to specify how and where the data are backed up and how to utilise data warehousing techniques to make the most updated data available to researchers, while retaining copies of previous versions of such data to ensure that results can be readily verified, corroborated, or modified, if required.



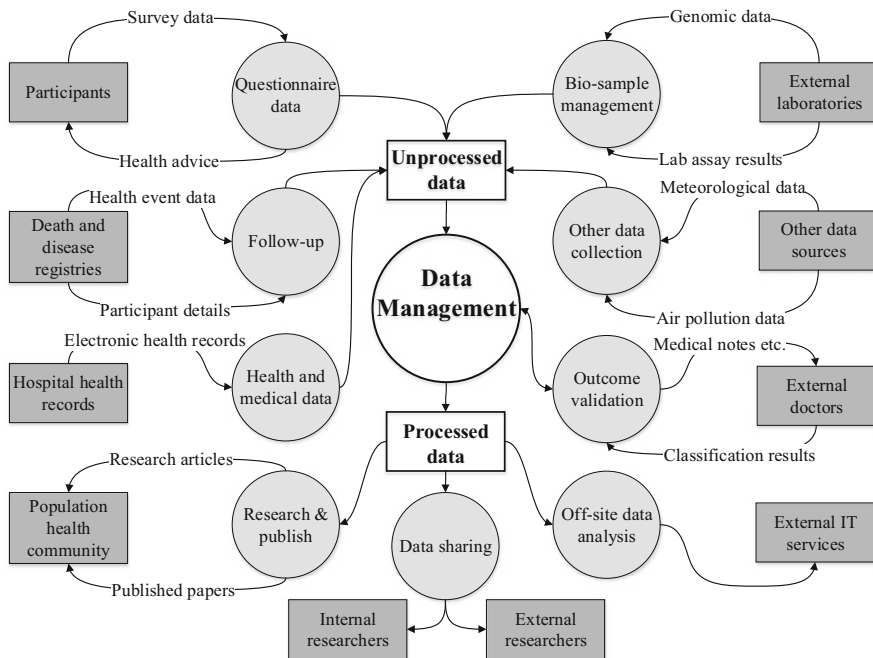


Fig. 8.2 Schematic representation of data management systems in CKB

### 8.3.2 Standard Operating Procedures

A Standard Operating Procedure (SOP) is a document that outlines the step-by-step instructions to conduct complex but routine operations involved in data management. Given the scope and complexity of tasks involved in management of data in large biobank studies, several distinct SOPs will be required, from the early planning stages to addressing emerging needs (Table 8.1). They clarify workflows, highlight areas requiring particular care, and ensure compliance with the appropriate regulations. Once developed, SOPs should be made readily available to the intended audience, including anyone who needs to understand the nitty-gritty of a particular process. Managing such documents with a version control system (e.g., Git) retains their history and development.

### 8.3.3 Naming Standards and Conventions

At the outset of a project, it is important to establish strict naming standards and conventions for data variables (including name, length, and format of each variable) and tables in which they are stored in the database. An efficient naming standard may

**Table 8.1** Examples of SOPs for data management in CKB

SOP	Intended audience	Contents
Questionnaire administration	Front-line staff	<ul style="list-style-type: none"> <li>• Advice for venue set-up</li> <li>• General question-asking technique</li> <li>• Answers to consent and data usage questions</li> <li>• Guidance for specific complex or personal questions</li> <li>• Handling uncooperative participants</li> </ul>
Data collection from death certificates	Data entry staff	<ul style="list-style-type: none"> <li>• Security policy for certificates and entered data</li> <li>• Linking certificates to study participants</li> <li>• Dealing with missing or unclear variables</li> <li>• Handling subsequent corrections</li> </ul>
Receiving health insurance data	Data management staff	<ul style="list-style-type: none"> <li>• Security policy for sensitive data</li> <li>• Determining and handling different file formats</li> <li>• Mapping fields supplied to fields requested</li> <li>• Checking for inclusion of all mandatory fields</li> <li>• Feeding issues back to data providers</li> </ul>
Integrating health insurance data	Data scientists	<ul style="list-style-type: none"> <li>• Security policy for sensitive data</li> <li>• Checking and importing files</li> <li>• Addressing data errors detected by automated validation</li> <li>• Addressing data warnings raised by automated checking</li> <li>• Feeding issues back to data managers/providers</li> </ul>
Responding to data requests	Data distribution staff	<ul style="list-style-type: none"> <li>• Verification of researcher and project</li> <li>• Checking that scope of request matches research goal</li> <li>• When and how to reject requests or request amendments</li> <li>• Producing and delivering datasets</li> <li>• Data documentation and FAQ for responding to queries</li> </ul>

afford an opportunity to derive useful information from the name of the variable, by including the source or category of the data variable within the variable name. For example, variables collected during the baseline and resurvey questionnaires can be stored in tables labelled ‘baseline\_questionnaire’ and ‘resurvey\_questionnaire’, respectively.

### 8.3.4 *Tools for Storing and Managing the Data*

Software purchasing choices, including Database Management Systems (DBMS), are covered in Chap. 7. For the data management tasks discussed in this chapter, the choice of database software is secondary, as the data structures required lend themselves well to a traditional Relational Database Management System (RDBMS) and any enterprise-level RDBMS will include the necessary tools (Foster and Godbole 2016). For example, PostgreSQL and MySQL are powerful, well supported, and free to download and install. SQL is the standard language for communicating with databases and is supported (with minor variations) by every RDBMS.

Coding techniques are beyond the scope of this chapter, but there are a number of core considerations for data manipulation. First, data should be added or edited not on an ad hoc basis, but by strict processes employing well-managed code. This code must be consistent and reliable, a goal best achieved by turning common tasks into applications, stored procedures, or functions—i.e., code units which can be rerun and reused with simple parameter changes. Second, version control (such as Git) is imperative, as a long-term study may generate questions about data gathered decades earlier. Finally, all code must be readable, which requires in-line comments, comprehensible variable and table names, and supplementary documentation.

Smaller studies may be tempted for simplicity's sake to store their data in spreadsheet software (e.g., Excel) but this approach has severe limitations and pitfalls. It does not scale well or allow for the power and scope of checking and analysis offered by a SQL-supporting database (Molinaro 2009). Excel, in particular, can introduce subtle but potentially irreparable data errors (Ziemann et al. 2016).

### 8.3.5 *Data Management Team*

Any study reliant on data requires a dedicated data management team. This team acts as an interface between developers, field staff, and analysts, and so must have excellent communication skills alongside specialised role-specific attributes. Some examples of data management roles are shown in Table 8.2, although, in practice,

**Table 8.2** Data management team specialisations

Role name	Role description
Head of data management	Manages all data management staff and day-to-day co-ordination of data management projects
Database architect	Designs database tables and other database schema elements
Database administrator (DBA)	Installs, configures, and maintains database systems, including managing backups
Data scientist	Integrates, transforms, and curates data into a form suitable for research
Database developer	Writes and maintains database software that satisfies specification

many staff will have combined roles or specialisations. Other functions, such as the role of the database administrator, may instead be centralised or outsourced.

## 8.4 Data Gathering and Validation

When gathering data, the key considerations are: (1) depth (incorporating every available source of useful data); (2) accuracy (though without necessarily rejecting imperfect or incomplete records); and (3) accessibility (requiring minimal subsequent cleaning and processing). Meeting these requirements requires an approach tailored to each data source and type. Since the objective is to store the data in electronic format, it should be gathered or entered electronically where possible and imported as soon as possible otherwise.

### 8.4.1 Data Sources

There are three main sources of data: (1) collected directly from participants by investigators (e.g., through interview, physical measurement, and sample assays); (2) collected about participants from primary sources (e.g., death certificates); and (3) collected indirectly from secondary sources (e.g., health insurance records). Input and validation methods depend on the source.

#### (a) *Data captured directly from participants*

When the participant is present, data collection can use strict acceptance criteria for answers, because any problematic results can be re-requested. Directly capturing data allows investigators to set limits on responses and require mandatory responses to key questions (Fig. 8.3). Further quality control procedures can be implemented at the time of data collection ('live') by checking the aggregate data on a daily or weekly basis, and immediately communicating with field staff about any problems identified.

#### (b) *Data captured directly from primary sources*

Apart from the participant themselves, the other important primary sources of data are those collected or generated elsewhere but linked directly to particular participants, such as occupational exposure records, death certificates, and hospital records. Depending on how assays are done, blood samples could also be considered in this category. As the data have been collected elsewhere, data capture should be more permissive (e.g., allowing missing variables), because it is generally preferable to have incomplete information rather than missing data (Fig. 8.4). However, some rules can still be enforced strictly, including validating IDs and dates. Checks should also be implemented to ensure that the data are being linked to the correct participants (see Sect. 8.6).

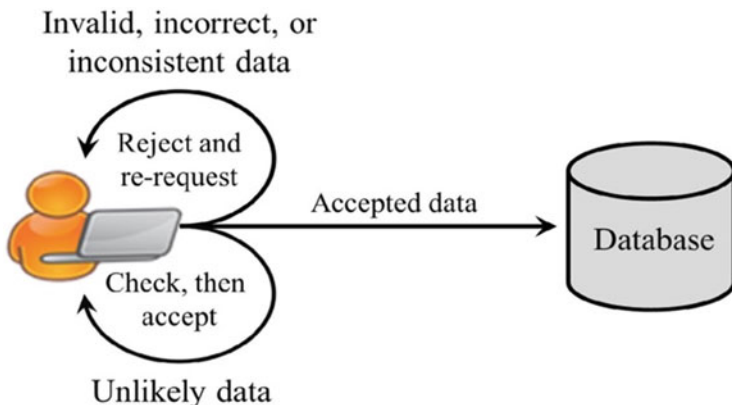


Fig. 8.3 Validation of directly collected data

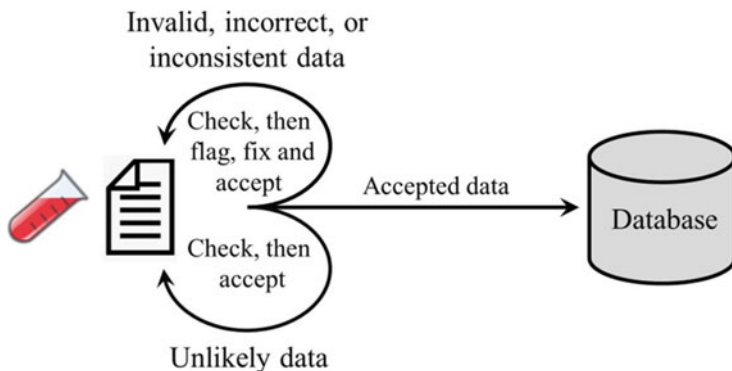
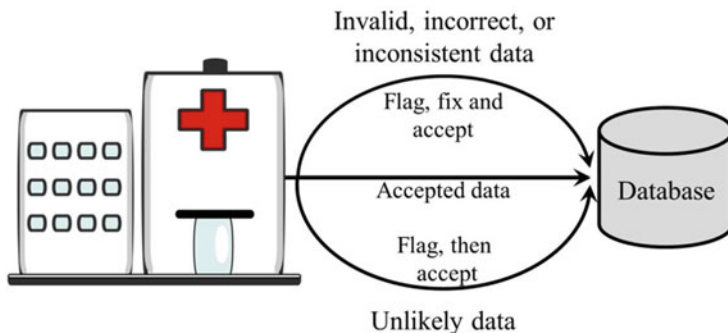


Fig. 8.4 Validation of data from a primary source

(c) *Data captured from secondary sources*

It is often necessary to collect and integrate data that have been collected from secondary sources by other organisations or investigators for other purposes (e.g., exposure to ambient air pollution, reimbursement data from health insurance agencies, or results of analyses of biological samples in routine hospital visits). Handling such data requires a more permissive approach: endeavouring to import everything, then check it automatically once it is in the database (Fig. 8.5). Decisions can be made about the handling of data problems on an individual or source level (e.g., counting bad dates as missing or giving them a default, or excluding a corrupted row in an otherwise valid dataset (see Sect. 8.5)). Particular care must be exercised to ensure that each record is linked with the correct participant (see Sect. 8.6).

Irrespective of the source of the data, it is essential to retain a copy in its original form. This enables investigators to verify accuracy, respond to questions, or resolve problems at a later date. Examples include: (1) scans of consent forms, disease



**Fig. 8.5** Validation of data from a secondary source

records, or death certificates; (2) photographs of medical records (e.g., test results); (3) data files received for import; and (4) original output files from biometric devices. Of course, these originals must be stored as securely as the data that was derived from them.

## 8.4.2 Data Types

Data typically fall into one of the following categories: (1) Boolean data; (2) categorical data; (3) date data; (4) continuous data; (5) free text data; and (6) image data (Kirkwood and Sterne 2003). Each has different issues to consider.

### (a) Boolean

Boolean variables permit only two possible values (e.g., True/False or Yes/No). Very few real-world variables are truly Boolean. Even when the goal is to force a dichotomy, it is generally better to permit more choice when collecting data and reassign the answers later. Where Boolean values *are* used, alternative possibilities can still be accommodated (e.g., allowing responses of ‘Yes’ or ‘No/Don’t know’).

### (b) Categorical

For multiple-choice questions, the possible responses must include: (1) clearly correct answers for the vast majority of respondents; (2) a manageable number of options; and (3) analytically useful categories of results. Allowing an ‘Other’ option (even where it may seem unnecessary) may help avoid spurious or missing answers. Integer values such as ‘number of siblings’ or ‘cigarettes smoked today’ can be treated in much the same way.

### (c) Date

Dates are easily validated, but consider the possibility of uncertainty, especially in areas which use different calendar systems (e.g., Chinese lunar calendar). Allowing

an answer of 'July 2015' might well be an improvement over forcing a specific date and thus getting either spurious accuracy or no value at all.

(d) *Continuous*

These are often physical measurements, such as height, weight, blood pressure, lung function, and blood lipid levels. Where possible, data should be collected electronically directly from the measuring device, though some devices (e.g., spirometers which measure exhalation volume every millisecond) can produce large volumes of data requiring special treatment for transfer and analysis. Managing numbers of arbitrary precision is surprisingly challenging (Goldberg 1991). It is advisable to choose a reasonable level of precision in advance based on measurement accuracy and analytical need, and truncate values to fit where necessary. Physical measurements often have physical limitations (such as standing height for some wheelchair users) so it is particularly important to allow for missing values, while documenting reasons where possible.

(e) *Free text*

Allowing respondents to input whatever they wish is very flexible, but results in data that are very difficult to validate and impossible to analyse without prior categorisation. Free text answers allow the inclusion of personally identifying (or otherwise inappropriate) data, and text entry may also require support for multiple character sets and require translation before use. Wherever possible, gather categorical data instead. A free-text pilot study can be used to establish appropriate categories if necessary.

(f) *Images*

Image data could include CT scans, ultrasound images, photographs of medical records, or scanned death certificates. Images have the same constraints for analysis as free text data, only more so. Almost any analyses of image data will be performed on variables extracted from such images. For example, carotid artery ultrasound images can be used to determine the presence or absence of plaques and the number of such plaques. Ideally, extraction of image data will be conducted using specialised software, though manual interpretation may be necessary. Irrespective of how the data are extracted, it is important to decide on the variables required and to design and test the extraction methodology, where possible, prior to commencing the data collection.

### 8.4.3 *Validation*

Data must be checked for validity as soon as possible, ideally at the point of collection. This process categorises data as invalid, incorrect, inconsistent, unlikely, accepted, or missing, with the boundaries of these categories depending on the specifics of each variable. As discussed above, the source of the data dictates how validation issues can be handled.

(a) *Invalid*

An invalid value could be text where a number is expected, or a date such as 30th February. Checking for such invalid data can be particularly challenging, because they generally cannot even be stored in a field designed for the expected data or data type. This can be minimised by validating the data at source wherever possible (see Chap. 7), but it is important to consider this possibility when handling data collected by others.

(b) *Incorrect*

A valid value may nevertheless be obviously incorrect, such as a standing height of 5 m, or a date of death later than the current calendar date. These types of error are more readily detected and processed than invalid data values, because at least they do not violate database rules about data types. Checks normally involve specifying upper and lower limits for each variable, though care must be taken to avoid excluding genuine outliers.

(c) *Inconsistent*

An inconsistent value is one that is incorrect *in context*. It is not impossible for participant to have smoked for 30 years or to be 25 years old, but it *is* impossible for both values to be true for the same individual. Devising consistency checks can be challenging, not least because errors may have several sources. A firm epidemiological and medical knowledge of the questions involved and their logical connections is essential. However, once the checks are determined, their implementation is generally simpler, such as setting limits on one variable (e.g., years spent smoking) that are derived from another variable (e.g., an upper limit of ‘years alive’). Error messages or other feedback about inconsistent values must be carefully worded to ensure that it is clear how and where they should be addressed.

(d) *Unlikely*

An unlikely value is one that is not impossible, but deserves additional checking. For example, though some individuals genuinely have a standing height of 2.5 m, that result is more likely to be a measurement or data entry error. Such issues are handled by requesting a manual double-check from the data gatherer. If confirmed, the value is accepted. Checks are generally a matter of specifying upper and lower limits for each variable, but assigning such limits requires some subjective judgement. There is a trade-off to be made between being too permissive and accepting dubious values, and requesting too many checks and thus undermining the value of each. More important variables may warrant stricter checks.

(e) *Accepted*

A value which does not meet any of the above criteria is accepted. It is important to appreciate that one cannot always be *certain* that data is correct, one can only limit the ways in which it can be erroneous.



(f) *Missing*

Missing values can be challenging, because their meaning and importance may depend on their context: a missing date of birth is a problem for most analyses, but a missing ‘cigarettes smoked per day’ is entirely appropriate for a non-smoker. In practice, for each variable, ‘missing’ falls into one of the above categories (impossible for some, acceptable for others.) A desire for completeness should not result in the rejection of an entire record for trivial omissions, or force users to enter ‘dummy’ answers to bypass validity checks. When gathering data, offer an explicit alternative of ‘unknown’ whenever appropriate (e.g., ‘birth weight’). This allows quality monitoring and subsequent analyses to differentiate these ‘genuine’ missing values from those arising from oversights or the removal of invalid answers.

## 8.5 Data Cleaning and Standardisation

The chief requirements for data management are to provide data that are reliable (accurate values), consistent (i.e., using uniform values throughout the records), and usable (in a research-ready form). The data cleaning and standardisation strategies required to attain these goals depend on the source and quality of the data.

### 8.5.1 *Handling Problem Data*

Despite best efforts to minimise the risk of such errors, problems within datasets invariably occur. There are several possible approaches to handling problem data, including exclusion, removal, correction, or flagging.

(a) *Exclusion*

The simplest approach is to exclude any record with a problem. However, exclusion risks losing valuable data, particularly if each record contains multiple mostly independent variables, where one mistake does not imply problems elsewhere. Exclusion may also introduce bias if the cause of the problem is non-random. For example, excluding participants with uncertain birth dates could disproportionately affect older people, or people from regions with less formalised record-keeping.

(b) *Removal*

A problematic value for a particular variable can be removed, leaving the remaining data intact. This is the best approach in most cases, though for some analyses it may be important to distinguish whether a value is missing for this reason or another (e.g., whether ‘date of death’ is unavailable because the participant has not died, or because it was removed for being invalid).

### (c) *Correction*

In some cases, it may be possible to infer a value to replace the problematic one. For example, if a questionnaire's completion date is missing, we could substitute the date that a blood sample was collected, if the two generally occurred on the same day. Again, it is important to ensure that this does not introduce bias into the results. For example, if blood glucose results are missing because the test cannot detect values below a certain threshold, it would clearly be inappropriate to replace those missing values with the average participant blood glucose. However, leaving those values as missing would also be misleading because we know a great deal about the true value, despite not having an exact number. It might be better to insert a result of zero or of the lowest detectable value, though neither is ideal. Whatever the decision, careful documentation is essential.

### (d) *Flagging*

It is tempting to simply flag any problems detected and leave individual analysts to decide how best to handle them. However, it can be particularly challenging to ensure that such flags are detected and understood before commencing analyses. It also duplicates work and risks inconsistent results from the same dataset. Wherever possible, one of the above approaches is preferable.

The key to any decision is to implement decisions consistently and automatically, and document them thoroughly and accessibly. This minimises the risk of errors and enhances confidence in the results.

## **8.5.2 Consistency Checks**

When collecting data, it is important to seek consistency between records and across different questions and/or measurements. This can be ensured by (1) standardising wording between questions and questionnaires; (2) offering the same sets of categorical answers wherever possible; (3) applying the same data validation rules; and (4) deploying the same measurement devices, consistently calibrated. Where changes are unavoidable they should be recorded in the database and documented for researchers.

## **8.5.3 Standardisation**

As discussed in Sect. 8.4.2, some data types are easier to analyse than others, with free text being one of the most challenging. However, most clinical systems are not designed with epidemiology in mind and often supply diagnoses as free text responses. Finding all instances of a particular disease in such data would require a lengthy and complicated keyword matching process unique to that disease. As

such, it is better to map text descriptions to standardised disease codes such as those provided in ICD-10, (WHO 2016) which are widely used and understood and can be used for unambiguous disease definitions. If necessary, bespoke software should be developed to automate and simplify disease standardisation processes. This process may involve: (1) importing the disease diagnosis text; (2) splitting any descriptions which contain multiple diagnoses (either automatically using punctuation, manually, or a combination of the two); (3) determining a disease of interest, based on frequency or analytical importance; (4) developing a set of keywords to identify this disease within a set of diagnoses; (5) having an expert manually assign an ICD-10 code to each diagnosis that matches those keywords; and finally (6) double-checking of all coding by a second expert, with a review process to resolve disagreements.

Standardisation is an iterative process, in which keyword choices are tested and refined using the remaining data. The goals are to detect any evidence of the disease of interest, even if this also means including some additional diseases. The matching typically allows rules such as ‘word1 AND word2’ and ‘word3 WITHOUT word4’. This process gradually develops a comprehensive dictionary of disease descriptions, capable of standardising even misspelled or abbreviated descriptions, which can be applied to subsequent batches of data. Novel descriptions will still have to be manually standardised as above, but an increasing percentage will already appear in the dictionary allowing them to be coded automatically.

## 8.6 Data Linkage and Integration

The primary unit of almost all biobank research is the participant, identified by a unique anonymous person ID. Researchers need to be confident that: (1) the ID records the correct participant and (2) records with different person IDs refer to different participants. Hence, reliable and accurate data linkage is very important in biobank studies.

### 8.6.1 Linkage Keys

Many types of linkage keys are available and this section outlines some of the most widely used and how they can be implemented in large biobank studies. In the CKB, for example, in which most data were collected electronically using bespoke software designed for the project, it was possible to mainly use deterministic (or exact) record linkage for matching information on participants across datasets. Deterministic linking can use a unique natural key, such as a government allocated National ID number, or a study generated unique ID, to determine if the records refer to the same participant. If unique identifiers are unreliable or unavailable for linking purposes, there are a number of other options available. Alternative approaches

**Table 8.3** Example of linkage keys used in CKB

Linkage key	Description	Unique	Links participants	Links events	Links treatment	Links samples
Baseline Study ID	Identifier of participant allocated by study at initial survey	Yes	✓			
Resurvey Study ID	Identifier of participant allocated by study at follow-up survey	Yes	✓			
National ID (NID)	Identifier of participant allocated by government	Yes				
Health insurance number	Number to identify claimant allocated by regional health insurance agency	No	✓	✓	✓	
Hospital number	Number to identify patient allocated by hospital	No	✓	✓	✓	
Cryovial ID	Identifier of cryovial with participant sample	Yes	✓			✓

may use probabilistic linking; additional details of other types of linkage techniques have been previously described elsewhere (Harron et al. 2016).

Table 8.3 provides examples of linkage keys from different data providers. In many cases, data will be linked using natural keys, which will then be replaced by internal study identifiers after confirming correct participant linkage. When a natural key is not unique, it can be modified by including additional variables to make such keys unique through the combination of variables.

In CKB, every participant was allocated a unique study identifier at baseline (e.g., K990000811). At subsequent surveys, the participant was allocated a new unique resurvey Study ID (e.g., K991000655), with the first two digits for study area (99) unchanged and the third digit indicating survey number ('0' for baseline and '1' for the first resurvey). The baseline Study ID and resurvey Study ID were then linked using a separate linkage table. The rationale for allowing two separate IDs is that, on rare occasions, a resurvey participant may inadvertently bypass the best processes and be matched to the wrong baseline Study ID (most likely their partner's ID). Correcting such an error by amending a linkage table is simple and can be easily audited. A similar technique is used for biological samples, where tubes are allocated a cryovial identifier rather than the Study ID of the participant (see Chap. 4), and linked back to a participant using a separate linkage table. Any ID ought to have a final check digit or code to detect data entry errors (e.g., a Verhoeff check digit, as described in Chap. 7).

### ***8.6.2 Identifying Discrepancies and Linkage Errors***

It is important to develop fully automated systems to link participants between multiple datasets, and check the validity of those linkages (Table 8.2). It is easy to assume that just because a record has a valid linkage key that the linkage is correct. However, it is possible that a source file collected externally for linkage to study participants has been incorrectly produced or that a code change has introduced an error. The risk of linkage errors is greater when natural keys are used, such as National IDs. Blindly trusting such linkage could lead, for example, to a woman being assigned her husband's medical record because his health insurance is in her name and hence linked to her National ID.

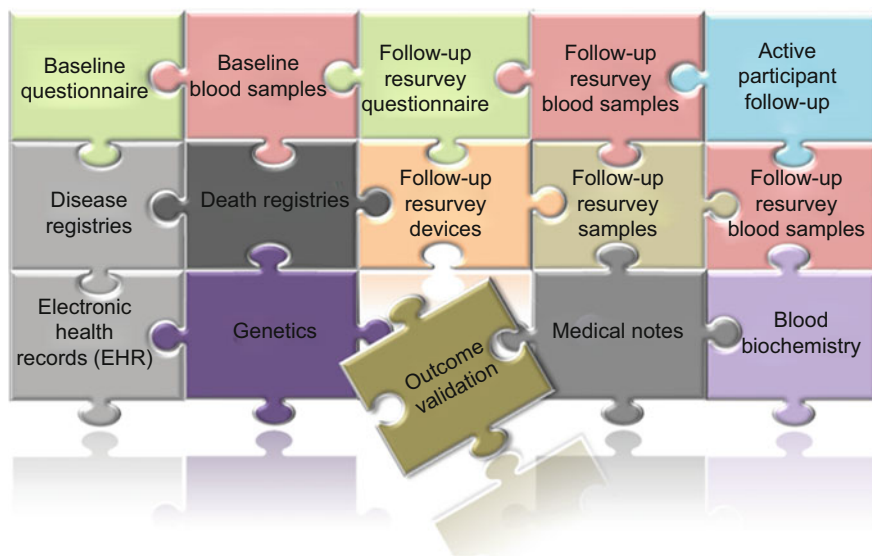
Typical checks conducted prior to linkage include: (1) comparing baseline survey personal information with health data personal information; (2) checking for anomalies such as events after death; (3) checking each participant's prevalence or incidence of disease between datasets; (4) checking gender-specific diseases; and (5) checking that regional health insurance links only to participants from the correct region.

Once any linkage discrepancies have been identified, it is the responsibility of the data management team to resolve any potential linkage errors. In many cases, it may be necessary to discard the problem records, because in general it would be preferable to include false negatives (missing data) than false positives (incorrectly attributed data) in the study.

### ***8.6.3 Integration of Data from Different Sources***

Using matching on linkage keys, as previously outlined, studies typically develop methodology and platforms (see Sect. 8.9) to enable collection (electronically wherever possible) and linkage of participant data from the following categories: (1) participant health data (e.g., baseline and resurvey questionnaires); (2) participant follow-up (e.g., death and disease registries and hospital admission records); and (3) participant population health data (e.g., administrative data and air pollution data) (Fig. 8.6).

Actively combining data from diverse sources enriches the detail and value of datasets and enables researchers to conduct high quality research. Therefore, it is important to develop an SOP for each source to describe how the data should be acquired, checked, and integrated. This can serve as the basis for a contract between the data provider and your study. While individual SOPs may differ, they should all address the following questions: (1) Have the files been processed or converted in any way? (2) Are there any supplementary data (e.g., images or raw waveforms)? (3) Have the data already been linked to participants or samples? If so, how? If not, what information is supplied to allow linkage to be performed? (4) Which device (if any) produced each record? (Both model number (the sort of machine) and serial



**Fig. 8.6** Schematic representation of linkage of different types of data in biobank studies

number (the specific machine) can help when trouble-shooting any data problems); (5) Do the data sources have a formal definition or API available? (6) How will the data arrive? (7) What do missing values signify? (8) What do duplicate records signify? (9) What processing and checks are required before the data are ready for release?

#### **8.6.4** *Updating Personal Information*

During the course of study, changes to participants' personal information are inevitable. These might be updates (e.g., to address), corrections (e.g., to date of birth), or less easily characterised (e.g., gender changes). It is important to have a consistent and well-documented strategy for handling such changes, particularly for core variables, such as dates of birth, where a change might move a participant between cohorts (e.g., age sub-cohort) or prompt questions about previously acceptable questionnaire responses. Irrespective of which approach is used, the priority should be to retain the original data and keep a detailed record of any changes made.

In general, the most recent update should be treated as canonical. It is important to allow the study to constantly improve data quality. This will have the unexpected, but unavoidable, result that each release of the study data may have a slightly different number of participants as people move into or out of the study age range. Many analyses also need to account for participants who have moved out of a study region and are therefore deemed to be 'lost to follow-up'.

Finally, every study must allow participants to choose to withdraw entirely. Unlike the changes discussed above, this sort of change needs to be applied retrospectively, and may necessitate the permanent deletion of data. Great care must be taken to respect the participant's wishes and all relevant data protection and retention legislation.

## 8.7 Data Aggregation

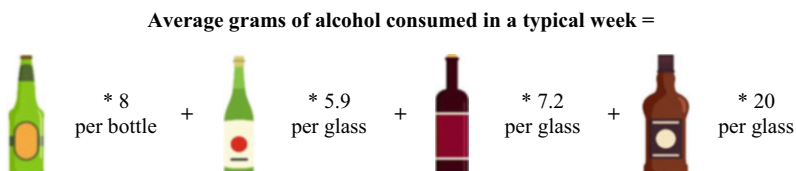
Data supplied for research should be available at the appropriate level of detail (i.e., neither too detailed, nor missing vital distinctions) and be consistently specified (meaningfully comparable with the rest of the study data, and ideally with other research databases). These requirements can be addressed by careful aggregation. Presenting data at multiple levels of detail empowers each analyst to choose how far to consolidate or subdivide each variable.

### 8.7.1 Raw Data

Analysts can simply be provided with the raw data (albeit checked and standardised as above). This has the advantage of allowing them to make their own decisions about relevance and groupings. The disadvantage is that it *forces* the analysts to make these decisions, duplicating effort and risking inconsistent results.

### 8.7.2 Derived Variables

For certain exposures with multiple questions involved (e.g., alcohol consumption, physical activities), it may be appropriate to construct a single derived variable that could provide a summary level of exposure (e.g., total alcohol amount consumed in grams/day, total MET-h in physical activity). These can then be included in the databases and made available to all data analysts.



**Fig. 8.7** Example derivation of an alcohol consumption variable used in CKB

For example, a researcher might take a detailed set of alcohol consumption questions and combine the answers into a single composite value (Fig. 8.7). Alcohol consumption could be further summarised into relevant categories such as: (1) never-regular drinker; (2) ex-regular drinker; (3) occasional or seasonal drinker; (4) monthly drinker; and (5) daily drinker. Once derived, such variables can be included in the database, allowing other researchers to use them without requiring every detail of the derivation, thus saving time and ensuring consistency between analyses. Furthermore, if new research suggests improvements to a formula, it can be updated in one place, and then the next data release can include the improved version for all other researchers to use.

### 8.7.3 Disease Outcomes

Unlike cause-specific mortality, electronic health records (e.g., episodes of hospitalisation, primary care data) can be extremely challenging to handle and analyse. Even after cleaning and standardisation, the results are derived from multiple records for each participant, covering different periods, and potentially containing different variables, duplicate events, or even conflicting information (a hospital admission dated after a death, for example). When combining records from multiple sources, all the challenges are multiplied. The goal is to transform data from these disparate sources into something that can be readily analysed, commonly: (1) one row per participant (using participant ID); (2) date of diagnosis; and (3) presence or absence of disease (requiring disease classification). An efficient and widely used approach to disease classification is ICD-10 codes, although other clinical coding systems are available (e.g., ICD-9, OPCS-4, Read v2, and Read v3 are all also used by UK Biobank).

**Table 8.4** Example of events for an individual participant from multiple sources in CKB

Participant ID	Source	Source variable	Diagnosis date	ICD-10 code (disease name)
990000811	Disease reporting	Stroke report	02-Jan-2009	I63 (ischaemic stroke)
990000811	Health Insurance	Admission diagnosis	05-Jan-2009	I63 (ischaemic stroke)
990000811	Health Insurance	Discharge diagnosis	07-Jan-2009	I61.1 (haemorrhagic stroke, cortical)
990000811	Disease reporting	Stroke report	18-Mar-2010	I61 (haemorrhagic stroke)
990000811	Death registry	Underlying Cause (1a)	01-Jun-2010	I21 (heart attack)
990000811	Health Insurance	Discharge diagnosis	05-Jul-2010	I10 (hypertension)



Some disease outcomes (especially deaths) may involve more than one diagnosis (e.g., underlying and contributing causes) and therefore may include more than one event on a single date. Combining variables and events from different sources might reveal inconsistencies, such as events that overlap or occur after death (see examples in Table 8.4). Such problems can be resolved by establishing a hierarchy of trust, removing data from less reliable sources where higher quality data are superior to lower quality data. It can be difficult to decide on whether events are duplicates or not. This method is designed to address that question only when absolutely necessary.

### 8.7.4 Applying Endpoint Definitions

ICD-10 includes several thousand disease codes. Depending on the number and range of disease outcomes captured, the study may simply supply individual ICD-10 codes for researchers to use, and/or generate common disease endpoint categories of specific interests to facilitate data analyses. Endpoints will vary between analyses, but standardised diagnoses make it easy to be flexible.

Once the endpoint definitions have been applied to each event, they can be aggregated for each participant (Table 8.5). A participant has an endpoint if they have at least one event which meets the endpoint criteria, and the date they first developed the endpoint is the date of the earliest such event. Where they do *not* have an endpoint, their censoring date is used instead, which might be their date of death, when they were lost to follow-up, or simply the end of the most recent follow-up period. Essentially, it represents the last date where we can be reasonably confident that they did not have that endpoint.

**Table 8.5** Example of classifying events into endpoints

Participant ID	Source	Source variable	Diagnosis date	ICD-10 code	Endpoints					
					01	02	03	04	05	06
990000811	Disease reporting	Stroke report	02-Jan-2009	I63	✓	✓	✗	✗	✗	✗
990000811	Health Insurance	Admission diagnosis	05-Jan-2009	I63	✓	✓	✗	✗	✗	✗
990000811	Health Insurance	Discharge diagnosis	07-Jan-2009	I61.1	✓	✓	✓	✓	✓	✓
990000811	Disease reporting	Stroke report	18-Mar-2010	I61	✓	✓	✓	✓	✓	✓
990000811	Death registry	Underlying Cause (1a)	01-Jun-2010	I21	✓	✓	✓	✓	✓	✓

*Endpoint definitions:* EP01: Any vascular disease (I00-I99); EP02: Any stroke (I60-I61 or I63-I64); EP03: Haemorrhagic stroke (I61); EP04: Death (as noted in death certificate); EP05: Second stroke (at least 28 days after first stroke diagnosis); EP06: Any cancer (Diagnosis: C00-C99)

The benefits of this approach to health record analysis include: (1) ease of adding new endpoints; (2) complex definitions still give simple results; and (3) ease of adding events from additional data sources.

## **8.8 Quality Control and Documentation**

It is important to measure and document the quality of all study data, including assessments of the accuracy, validity, and context of data collection. This may involve additional data gathering.

### **8.8.1 Reliability**

A finding is reliable if repeated measurements give consistent results. Replicated measurements in random samples are needed to assess the accuracy of any data collected. Reliability of questionnaire responses can be assessed using a quality control questionnaire: a small selection of the key questions from the baseline questionnaire. Administering this questionnaire to a randomly selected subset of study participants shortly after the initial questionnaire allows assessment of the reliability of the answers. Subsequent resurveys of questionnaires and physical measurements after 3 or 5 years allow assessment of the longer-term reliability or within-person variability of such measurements.

### **8.8.2 Validity**

The validity of self-reported data can be checked against data collected from other more reliable sources. For example, self-reported physical activity and sleeping patterns can be validated, to some extent, by collecting data from participants who have been provided with accelerometers. As a very different example, electronic health records collected by health insurance agencies for reimbursement purposes can be checked against primary medical records: Original hospital notes can be retrieved to verify the reported event and then, where appropriate, the evidence can be reviewed by specialists to adjudicate the reliability of the diagnosis (Fig. 8.8). Not only does this help to validate disease reports, but it also enables detailed sub-phenotyping of diagnoses (see Chap. 6).

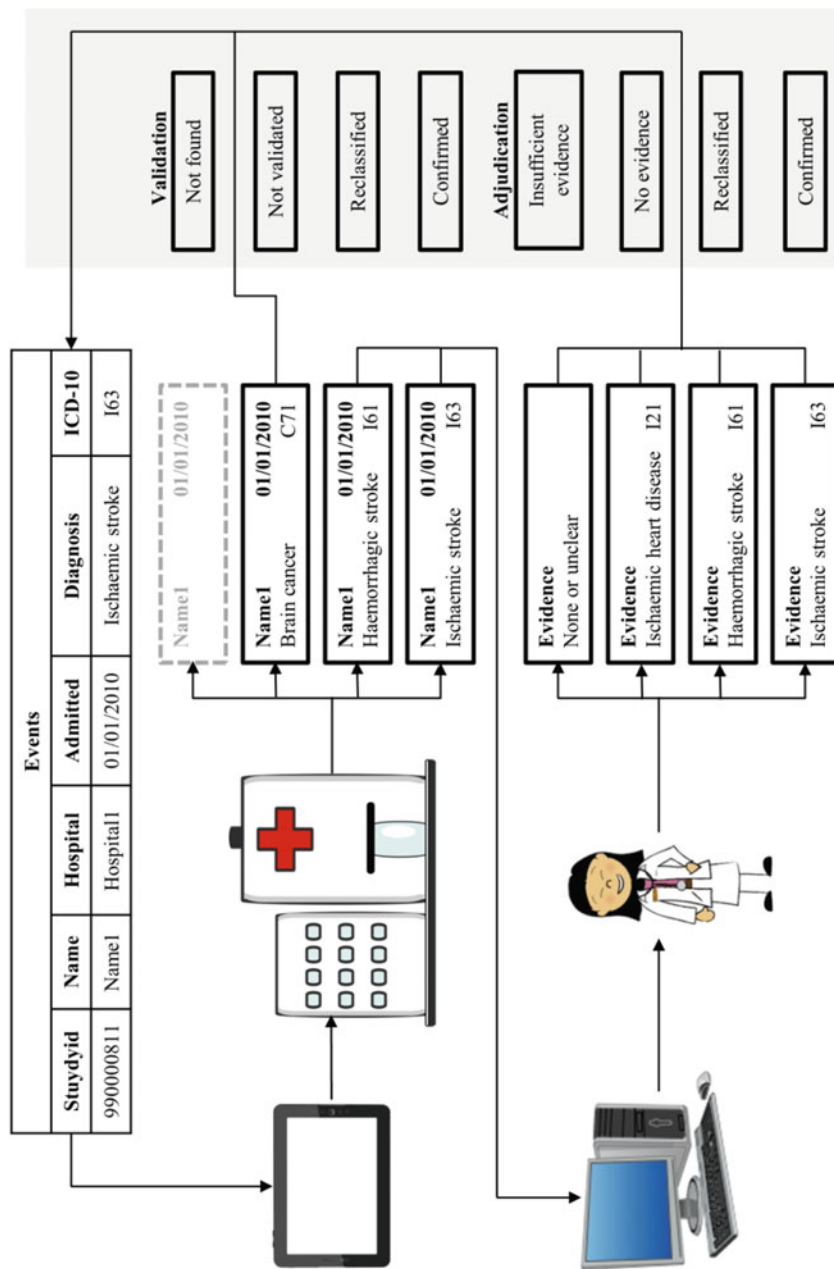


Fig. 8.8 Example workflow of disease validation system

### 8.8.3 Documentation

It is essential to document all data provided for analysis, even if it is intended solely for use within the organisation. This will include data collection methodology for each data source, and more detailed information about the variables available, including summary statistics to answer initial questions and highlight potential issues. For data to be shared with external researchers, documentation is even more important. The goal is to enable potential collaborators and users to understand the available data, and have confidence in the accuracy and validity of such data. The key objective should be to share all available data with *bone fide* researchers using appropriate study-specific data sharing platforms (e.g., <http://www.ckbiobank.org>, <http://www.ukbiobank.ac.uk>). Ongoing studies will periodically release new data for research purposes, and these should be accompanied by release notes providing details of new and updated data.

### 8.8.4 Metadata

Interpretation of any data requires prior knowledge of context, including knowing how, when, or why a variable was collected. Each variable supplied to a researcher should be accompanied by appropriate metadata. This should include the exact text question for questionnaire variables, the range of answers accepted, any validation or subsequent corrections that were performed, and codes for missing values.

## 8.9 Data Integration Platforms

Researchers require datasets that include: (1) up-to-date data (incorporating the latest data as soon as possible); (2) detailed data (including novel data sources as soon as they become available); (3) reliable data (using the latest improvements in linkage and data quality); (4) static data (resistant to change); and (5) reproducible data (can be recreated and extended, even long after the original dataset was delivered). These requirements demand a well-designed platform primarily administered by IT professionals, which can continuously collect, process, transform, and integrate data from multiple sources. Such platforms should implement the capability to produce and retain static copies of the data at a given calendar date (referred to as called ‘snapshots’), with appropriate and recorded censor dates.

### ***8.9.1 Development and Implementation***

A detailed data management plan is a fundamental prerequisite to the development and implementation of a data integration platform (see Sect. 8.3.1). Each study will inevitably have a number of defined tasks, requiring separate database environments with different functions (Fig. 8.9). Access to these environments must be carefully managed. It is critical that *all* access to *any* data is restricted as much as possible, with direct access to personal information being the most restricted.

The choice of database software is secondary, as discussed in Sect. 8.3.4. The hardware requirements for hosting such data may vary significantly depending on the size and complexity of the data, ranging from a standard server (with a storage capacity of a few TBs) to supercomputer processing facilities and cloud-based storage providers with the capacity to store many PBs of data (see Chap. 7).

### ***8.9.2 Live Environment***

The live environment is where data are accessed and updated in real time, and should be only accessible through applications within strict and well-defined limitations. In the case of CKB, this is where all operational data such as sample tracking and long-term follow-up of participants (including updates from the death and disease registries) arrive continually. Access to and management of this environment is the responsibility of the IT Management team. It is important to note that data stored in a live database will contain personal data, so access should be highly restricted through applications and only granted to those personnel whose roles require it. For example, a database administrator responsible for investigating issues with applications may need access to the underlying data. A developer working on data collection applications, even ones intended for use in the live environment, should typically be working on anonymised test data instead.

### ***8.9.3 Integration Environment***

The integration environment is typically used for importing, integrating and linking data sources (including the live data), combining them into a single database. Data arrive in diverse formats, generally as a file or collection of files, and are extracted, transformed into a common format, and loaded into database tables within the integration environment. From there, the data are cleaned, linked, and standardised, before the final process of generating outcomes and variables for each participant can be performed. Access to and management of this environment is the responsibility of the data management team.

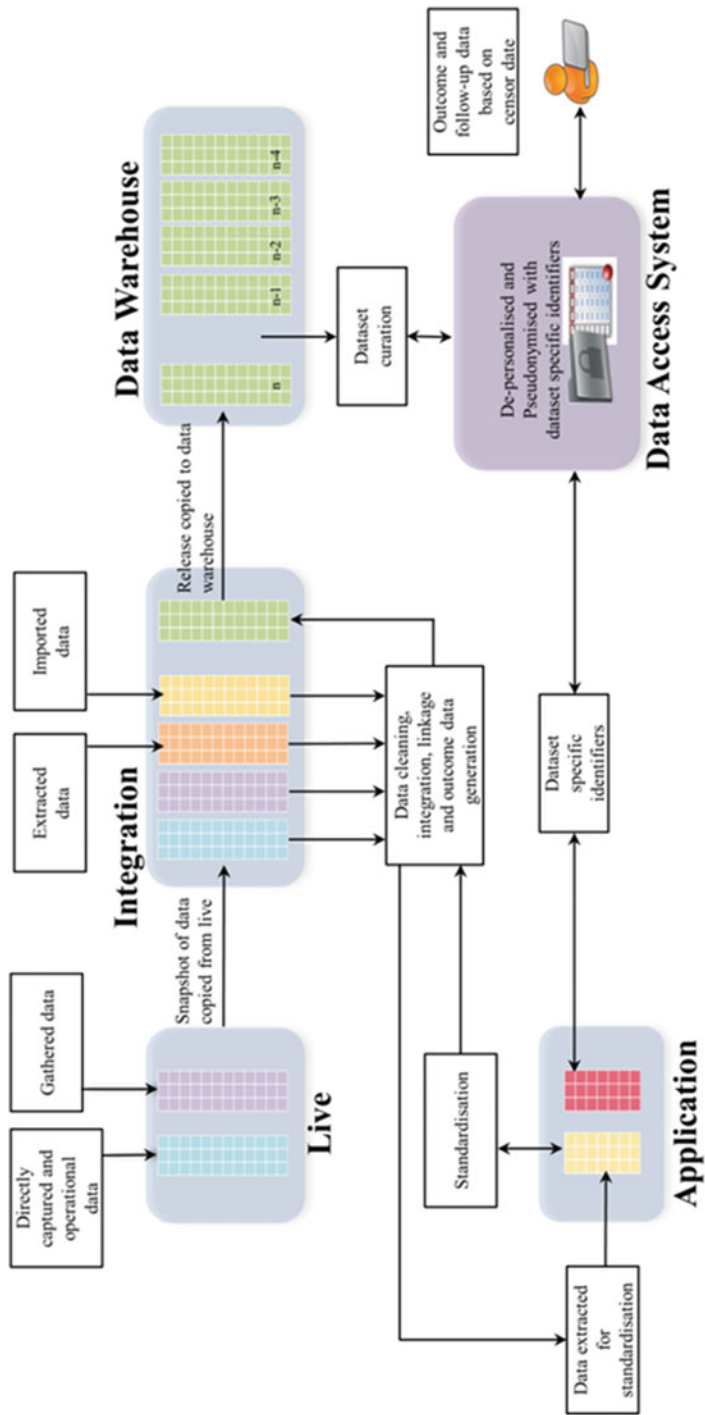


Fig. 8.9 Schematic representation of a data integration platform in CKB

### **8.9.4 *Application Environment***

A large biobank study will need to develop specific applications for use by the data management team. Some may be for managing data quality (e.g., standardising hospital names and disease descriptions), others for managing data delivery (e.g., generating Concealed Study Ids [CSIDs] to replace direct identifiers for research datasets). These applications are hosted in an environment that has all the disciplines of ‘live’, but under the responsibility of the data management team.

### **8.9.5 *Development Environment***

Whenever any data management applications or processes are changed, or any data from a new source are integrated, it is imperative that the results are tested thoroughly before being utilised. The development environment is a copy (or copies) of the database, to be used for coding and testing. As these tasks often involve changes to the data, ideally each developer will have their own copy of the database which they can backup, restore, and refresh independently. Access to and management of this environment is the responsibility of the data management team.

### **8.9.6 *Data Warehouse Environment***

The day-to-day work of the data management team culminates in regular releases of research-ready databases (‘snapshots’), typically on an annual basis. Each new snapshot will include: (1) the latest follow-up events; (2) any other newly integrated data; and (3) any corrections or improvements in linkage and data quality. This environment will *not* include any personal data, to minimise the risk that such data is shared inappropriately. All such releases are copied to the data warehouse environment where they will be retained in a read-only state for research and fieldwork. All requests for analysis datasets are served from these releases, generally the most recent one. Access to and management of the data warehouse environment is the responsibility of the data management team.

## **8.10 Governance and Access to Data**

Researchers require assurance that their data are ethically and legally appropriate to use. Compliance with appropriate ethical and legal regulations is also of fundamental importance to both biobank managers and participants. It is a key responsibility of the data management team to protect the integrity of the study, the privacy of study

participants, and the ethical obligations of the researchers. Safeguards must be maintained to ensure the anonymity and confidentiality of participants' data (Arbuckle and El Emam 2013).

Researchers must enter into a legal agreement not to make any attempt to de-anonymise participants. Data provided to researchers will not contain any personally identifiable variables and every dataset provided will be 'anonymised' with a uniquely encrypted identifier for participants. There are several steps that should be established before data are made available to researchers and these are summarised in Fig. 8.10.

### 8.10.1 Researcher Registration

All researchers seeking data need to be registered as *bona fide* researchers of the study. Registration is typically recorded using a public web-based interface (UK Biobank 2011). Each application must be individually reviewed and approved by members of the study Data Access Committee.

Researchers should be employees of a recognised academic institution, health service organisation, or charitable research organisation, with appropriate experience in conduct of health-related research. Many studies (e.g., UK Biobank) may also allow researchers from commercial organisations to access the data. Researchers should be able to demonstrate, through their peer-reviewed publications in the area of interest, their ability to carry out the proposed study. Their organisation should have formal policies and procedures to comply with any legal, ethical, or data

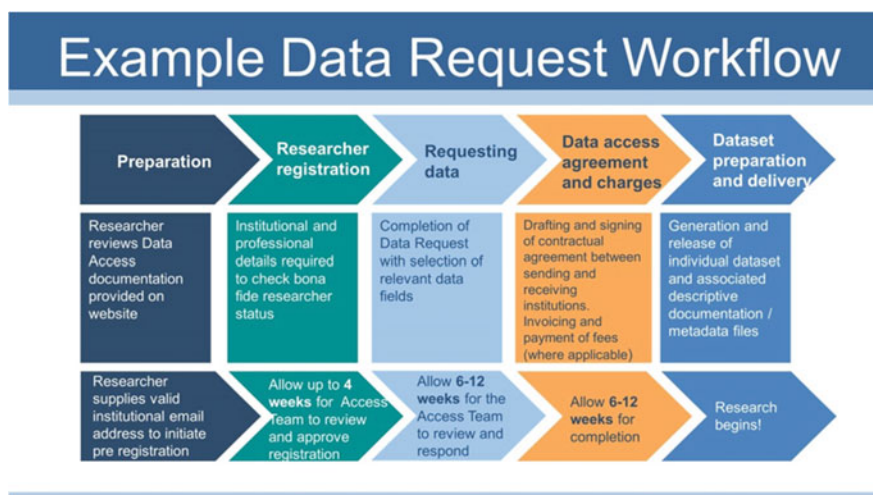


Fig. 8.10 Schematic representation of a data request workflow in biobank studies



protection constraints, and to ensure that the dataset is stored securely and used responsibly.

It is possible that a study has internal researchers within the study establishment as well as a commitment to supply data to ‘external’ researchers. It is strongly recommended that the process of registration and of requesting data is followed consistently by all categories of researchers.

### ***8.10.2 Requesting Data***

Researchers want to have the ability to: (1) learn what data is available to them; (2) request data easily and receive it quickly and securely; (3) add additional variables to previous requests; (4) update previous requests to reflect the latest data or continue to use older data; and (5) choose the format of their dataset. A data delivery system must address these requirements and, as far as possible, should automate them. Once a researcher has been approved, they will receive confirmation and a permanent registration ID. They can then be given visibility of the data that they can request using a ‘data showcase’. The data showcase must only display summary information about variables and not information at the participant level. The researcher can initiate a request using an online form and select the variables and categories of data that they would like to receive in their dataset. Once the researcher has completed their request they can submit it, at which point the request will be saved and the data access team notified. All data requests are typically reviewed by members of the study Data Access Committee and either the request will be approved or feedback with requests for modification will be communicated to researchers. Such responses might be an outright rejection, but are much more likely to suggest modifications and restrictions to reduce unduly large data requests. Once a request has been approved, the researcher will be notified, and a data access agreement will be prepared to control the use of the data.

### ***8.10.3 Data Access Agreement***

Before any data are released to researchers, a Data Access Governance agreement (DAG) should be agreed and signed between the biobank establishment and the researchers’ host institution. Such an agreement will specify the permitted use of the data. One important component includes a statement that the researcher agrees that they will only use the data for the permitted use and will not attempt to reverse engineer data or de-anonymise individuals.

To cover costs associated with the application process and the preparation of the dataset for each data request, it is standard practice to publish access costs to the resource so that applicants are aware that they will have to pay for a successful data request. If the request has been submitted by an internal researcher within the study

institute then, depending on the study's policies, the step for a DAG or for charges for the dataset can be bypassed.

#### ***8.10.4 Preparation and Sharing of Study Data***

Once a request has been formally approved by the Data Access Committee, the data access system will be notified and will initiate construction of a customised dataset. Curation of such datasets should be almost fully automated to retain consistency and reproducibility of datasets and avoid the risks of unintentional disclosure of direct or personally identifiable variables.

Access to all underlying and raw data must be restricted via applications or to only those personnel with exclusive privileges for accessing the data such as database administrators or data analysts, whose role is to curate the study data. Each dataset will be individually prepared using the selected variables and outcomes that the researcher has requested. The researcher will not have any direct access to the data via the web-based interface.

There are a number of stages in preparing a dataset before it can be released to the researcher (see Sect. 8.10) and careful processes must be put in place, for example, to avoid: (1) supplying variables that are classified as identifying and (2) unintentional disclosure such as non-identifying attributes containing identifying information (e.g., home address). In developing processes to curate researcher datasets, data managers must always consider that: (1) privacy and security must be preserved and (2) protecting privacy of genomic information is a major challenge. Once a dataset has been curated and is ready to be released to the researcher, the researcher will be notified and, through the web-based interface, will be permitted to logon and download their encrypted custom-built dataset.

#### ***8.10.5 Curation of Research Datasets***

There are legal and ethical restrictions on the types of data that should be released for research. Legal restrictions prevent release of direct identifiers, and safeguards must be maintained to ensure the anonymity and confidentiality of participants' data. Any individual variables that enable identification in isolation should never be included in a data request. Examples of direct identifiers include: name; address; date of birth; telephone numbers; email addresses; national ID numbers; medical record numbers; biometric identifiers (including finger and voice prints); or full-face photographic images.

Other variables may be classed as indirect or non-identifiers. Indirect identifiers are variables that could enable de-anonymisation when used in conjunction with other attributes held either internally or externally in a dataset. In the contemporary era with advanced computing capabilities, there is always a risk about what level of

granularity should be provided for data released; in addition, there should be awareness of the potential for rare or unusual diseases to indirectly identify a participant. For example, while height would not normally be considered an identifier of any kind, it could be for someone who is extremely tall. Non-identifiers are those variables that provide an extremely low risk of de-anonymisation and can be safely released. Classification of what constitutes a direct or indirect identifier is frequently a matter of human judgement, although it may be prudent to discuss any concerns with the relevant legal and ethical department, who can provide predefined classifications of common attributes.

All approved datasets should be pseudo-anonymised, replacing even internal study IDs with cryptographically secure dataset-specific identifiers. This will discourage unauthorised ‘joining’ of any two supplied datasets to create a broader dataset outside of the scope of the agreed research proposal.

### ***8.10.6 Delivery of Approved Datasets***

The sharing and delivery of study datasets to researchers is usually conducted using web-based interfaces including individual logon credentials. Such interfaces typically use secure network locations for internal researchers or encrypted internet delivery systems for external researchers. As an example, when datasets are ready to download, researchers receive an email with half of an encryption key. The second half of the encryption key is obtained by logging onto the web-based interface using their registration ID. The encrypted datasets can then be downloaded and unencrypted by combining the two parts of the encryption keys. Each dataset is delivered with an associated set of metadata files describing the data. The data access management system records and audits all datasets that have been prepared and delivered.

## **8.11 Summary**

The size, diversity, complexity, and potential sensitivity of contemporary large biobank studies demand secure and reliable data management frameworks. This chapter describes some of the key concepts, principles, and practical procedures involved in planning, designing, developing, and implementing these frameworks. Many of the procedures and practical examples illustrated were based on CKB, which has been accumulating large and multi-dimensional data from many sources since 2006. It is expected that technological development in genomics will soon enable whole genome sequencing of all study participants, along with multi-omics assays of many thousands of proteins and small molecules. These will generate unprecedentedly large and complex datasets, over and above those collected and generated using more conventional approaches. Advances in data science

technology, such as machine learning and cloud computing, may permit greater capacity, efficiency, flexibility, and novel solutions for managing and analysing such datasets, in ways that differ importantly from those described in this chapter. However, the fundamental principles will not change.

## References

- Arbuckle L, El Emam K. Anonymizing health data – case studies and methods to get you started. Newton: O'Reilly Media; 2013.
- Foster EC, Godbole S. Database systems - a pragmatic approach. New York: Apress; 2016.
- Goldberg D. What every computer scientist should know about floating-point arithmetic. *ACM Comput Surv.* 1991;23(1). <https://dl.acm.org/doi/pdf/10.1145/103162.103163>
- Harron K, Goldstein H, Dibben C. Methodological developments in data linkage. London: Wiley; 2016.
- Kirkwood BR, Sterne JAC. Essential medical statistics. 2nd ed. Hoboken: Wiley-Blackwell; 2003.
- Molinaro A. SQL cookbook – query solutions and techniques for database developers. Newton: O'Reilly Media; 2009.
- UK Biobank Limited. UK Biobank: Access procedures November 2011. 2011. Available from <http://www.ukbiobank.ac.uk/wp-content/uploads/2012/09/Access-Procedures-2011.pdf>
- World Health Organisation International Statistical Classification of Diseases and Related Health Problems 10th Revision. 2016. Available from: <https://icd.who.int/browse10/2016/en>
- Ziemann M, Eren Y, El-Osta A. Gene name errors are widespread in the scientific literature. *Genome Biol.* 2016;17:177.