# PROTOCOL FOR THE UK BIOBANK

A study of genes, environment and health

**14 February 2002**

## TABLE OF CONTENTS

**TABLES**

**FIGURES**

# 1. SUMMARY


## 1.1 Background

Despite the longstanding awareness that disease risk relates to the factors an individual is exposed to during their lifetime and to their genetic susceptibility (and to chance), a clear picture of the combined effects of genotype and exposure on disease risk is yet to emerge. Studies to date have often been characterised by small sample size (yielding inconsistent and statistically unreliable results), incomplete or inadequate measures of exposure, lack of formal statistical testing of effects and use of a retrospective case-control design. In order to translate recent advances in genetics and epidemiology into reliable information of direct clinical, aetiological and public health relevance, there is a pressing need for comprehensive, large scale, reliable data on the combined effects of genotype and exposure on the risk of disease. A large population-based prospective study would provide such data on a wide range of conditions, exposures and genotypes and would constitute an important resource for future research. Due to the unique combination of a large heterogeneous population and a centralised National Health Service, the United Kingdom is in an ideal position to conduct such a study.

## 1.2 Aims

The main aim of the study is to investigate the separate and combined effects of genetic and environmental factors (including lifestyle, physiological and environmental exposures) on the risk of common multifactorial diseases of adult life.

## 1.3 Methods

The investigation will take the form of a prospective investigation (known as "the UK Biobank") involving at least 500,000 men and women aged 45-69 from the general population of the United Kingdom.

***Recruitment***: People registered with participating general practices will be asked to join the study by completing a self-administered questionnaire, attending an interview and examination by a research nurse, giving a blood sample and providing written consent for participation and follow-up. Blood samples will be processed and stored to allow retrieval for nested case-control studies.

***Follow-up***: Participants will be flagged through the Office of National Statistics, providing routine follow-up data regarding cause-specific mortality and cancer incidence. Data regarding incident morbidity will be obtained through regular follow-up via hospitalisation data and general practice records, with confirmation of diagnoses using standard criteria. Every two years a subset of around 2,000 participants will be resurveyed to allow correction for regression dilution and the entire cohort will be resurveyed by postal questionnaire at 5 years to update exposure data and to ascertain self-reported incident morbidity. While follow-up is already feasible using existing hospital and general practice resources, it will also make use of NHS information systems where appropriate and the study design will be adapted as new information resources become available.

***Analysis, numbers of events and power****:* The main means of assessing the combined effects of genotype and exposure on the risk of disease will be through a series of case-control studies nested within the cohort. These will provide information regarding the risk of disease associated with exposures of interest, for individuals with and without a particular genotype, and will also give information regarding the effect of genotype on the risk of disease in people with and without particular exposures. Evidence of formal statistical interaction between genotype and exposure will also be examined. Studies based on continuous phenotypic outcomes are also envisaged within the cohort.

For conditions yielding around 5,000 or more incident cases (e.g. diabetes mellitus, ischaemic heart disease mortality, myocardial infarction, colorectal cancer, breast cancer), the study will have the power to detect minimum relative risks of disease associated with exposure (within the genotypes of interest) of 1.5 and an interaction ratio of around 1.4, for exposures and genotypes present in around 20 to 80% of participants (at 95% power and 0.1% significance). For the same exposure, genotype, power and significance values and outcomes with around 1,000-2,000 events (e.g. rheumatoid arthritis, Parkinson's disease, hip fracture, ovarian cancer, bladder cancer, etc), the study should be in a position to detect relative risks of disease associated with a particular genotype and/or exposure occurring of around 1.8-2.0 and an accompanying interaction ratio of 1.7-2.0.

## 1.4 Expected outcomes and contribution to knowledge

As the largest and most comprehensive prospective study with biological samples in the world, the UK Biobank is expected to contribute substantially to international knowledge regarding the combined effects of genotype and exposure on the risk of disease. Its design means that the study will provide a structure and resources for future research, and will enable researchers to address current and unforeseen scientific questions.

## 2. DESCRIPTION OF THE PROJECT

### 2.1 Purpose

#### 2.1.1 Nature of the proposal and terminology

This document is the protocol for the large scale investigation of combined genotype/exposure effects on disease risk (known as "the UK Biobank") and provides a broad description of its background, general principles, structure and methods. It is an application for the first five years of support for the initiative, however it includes information on aspects of the project occurring after this time period, for completeness.

While the proposal focuses on the scientific justification for the setting up of a large population-based cohort of men and women in the UK and subsequent nested case-control studies, it should be noted that the UK Biobank is designed to provide a broad and comprehensive framework for future research into the combined effects of genotype and exposure to various factors. By collecting and storing biological samples and detailed data on exposure on at least 500,000 individuals aged 45-69 the project will constitute a national research resource available to the scientific community for a large number of studies. At this stage, the detailed studies envisaged within the cohort include nested case-control and cross-sectional studies (outlined in this document; including studies incorporating biochemical, proteomic and metabonomic measures), additional investigations of intermediate phenotypes and family based studies (support for these studies is not being requested here; see section 2.3.14). However, the framework is designed so as to minimise restrictive underlying assumptions about genotype/exposure/phenotype/outcome relationships, to allow for the range of novel study designs and research needs which are likely to emerge throughout the post-genome period. This protocol outlines, in principle, the main baseline and outcome measures. The process of more detailed development of the study instruments and infrastructure is given in section 3.

Throughout this document, the term "exposure" will be used to refer to factors other than an individual's genotype which may be related to the risk of various outcomes. These include demographic, environmental, lifestyle, reproductive, medical and physiological factors and are sometimes referred to as "environmental" factors. Although the term "gene-environment interaction" is sometimes used to refer to the combined effects of genotype and exposure on outcome, the use of the term "interaction" in this document is restricted to references to formal statistical interaction between genotype and exposure (i.e. departure from a multiplicative model; see section 2.3.11).

#### 2.1.2 Rationale

It has long been recognised that an individual's risk of disease relates to three broad areas: their exposure to various factors, their personal susceptibility to disease and to chance. Despite the long standing awareness of these aspects of disease risk, the ability to investigate the combined effects of exposure and genotype on the risk of disease has, until recently, been limited and a clear picture of these relationships is yet to emerge. A full understanding of the combined effect of genotype and environmental exposure requires the accurate quantification of the risk of disease associated with exposures of interest, for individuals with and without a particular genotype, and quantification of the effect of genotype on the risk of disease in people with and without particular exposures, on a large enough scale and with appropriate statistical methods to take the effect of chance into account. Examination of evidence for statistical interaction between genotype and exposure is also of importance.

Distinguished commentators have pointed out that taking advantage of the opportunities arising from the rapid developments in genetic knowledge and technology may well represent the next major advances in biology and medicine.[1,2] Bio-medical science has therefore reached the stage where large-scale population-based studies incorporating information on genetics and health are not only feasible, but are also the appropriate next step in translating recent advances, such as the mapping of the human genome, into knowledge of direct clinical and public health relevance. At the same time, progress in information technology has improved the feasibility and cost-effectiveness of conducting large-scale epidemiological studies. A unique and timely opportunity therefore exists for the setting up of a large study incorporating information on genetic factors and an individual's health and exposure history. Due to the unique combination of a large population and a centralised National Health Service, the United Kingdom is in an ideal position to conduct such a study.

The large scale of the UK Biobank means that it will be ideally placed to confirm or refute existing hypotheses regarding various exposure/genotype/outcome relationships which remain uncertain due to the constraints of the data which are currently available (see below). The study will also be able to generate and investigate novel hypotheses regarding combined exposure and genetic effects. The availability of prospectively gathered and appropriately stored biological material on all study participants, along with comprehensive measures of exposure and physical characteristics and careful ascertainment of disease endpoints will permit investigation of a wide range of exposure/genotype/outcome relationships. Reliable quantification of such relationships is the main goal. As well as being able to investigate the broad effects of genotype and exposure the study has the potential to answer many important questions relating to pharmacogenetics. In particular, the study will be able to examine whether or not the risk of adverse events relating to use of certain medications varies according to an individual's genotype.

The initiative will target common conditions which are responsible for high levels of morbidity and mortality in the general population, including cardiovascular, metabolic, musculoskeletal and neuropsychiatric conditions, and cancer (see section 2.3.11.3). Improved means of preventing, screening for and treating these conditions arising from the UK Biobank will have far reaching implications for the health of the public and the health of individuals.

Insights gained from the proposed study will therefore have important implications for a number of aspects of health and clinical practice:

- The improved understanding of disease aetiology is likely to have benefits across the board; from the provision of information relevant to basic science to the development of new means of preventing and treating disease.
- The more precise identification of individuals at increased risk of disease through both exposure and genotype will allow improved targeting of various interventions.
- Information regarding combined drug/genotype effects will help to minimise the adverse effects of pharmaceutical agents and facilitate the development of new treatments.
- The design of the study also means that the data contained within it can be used to address questions of future scientific and public health relevance which have not yet come to light.

### 2.1.3 Aims and objectives

The main aim of the study is to investigate the separate and combined effects of genetic and environmental factors (including lifestyle, physiological and environmental exposures) on the risk of common multifactorial diseases of adult life.

### 2.1.4 Example research hypotheses

The proposed study will be able to address a large number of existing hypotheses regarding the combined effects of genotype and exposure on a range of important health outcomes. In practice, the hypotheses to be investigated within the cohort will be defined by researchers who are successful in applying to use data from the study, once recruitment and follow-up are sufficiently complete (see section 2.3.1).

The number of hypotheses able to be addressed, and that fact that many new hypotheses are likely to emerge over the course of the study, means that a comprehensive list cannot be provided here. Examples of important current hypotheses relating to common exposures, polymorphisms and outcomes which will be able to be investigated with a relatively high degree of reliability (see section 2.3.11, below) in the study include:

1. That cigarette smoking and polymorphisms in genes such as those coding for endothelial nitric oxide synthase and Apolipoprotein E4 have independent and combined effects on the risk of ischaemic heart disease;
2. That plasma levels of homocysteine and polymorphisms in the gene coding for Apolipoprotein E4 have independent and combined effects on the risk of ischaemic stroke;
3. That adiposity and polymorphisms in genes relating to aspects of metabolism (e.g. the PPARG gene) have independent and combined effects on the risk of diabetes mellitus;
4. That alcohol consumption and smoking and polymorphisms in the gene coding for Apolipoprotein E4 have independent and combined effects on the risk of dementia;
5. That exposure to infections (such as retroviruses) and polymorphisms in the HLA DRB1 locus have independent and combined effects on the risk of rheumatoid arthritis;
6. That use of exogenous hormones and polymorphisms in the BRCA1 and BRCA2 genes have independent and combined effects on the risk of breast cancer;
7. That endogenous testosterone levels and polymorphisms in the gene coding for the androgen receptor have independent and combined effects on the risk of prostate cancer;
8. That the consumption of meat and polymorphisms in the gene coding for N-acetyl transferase 2 have independent and combined effects on the risk of colon cancer;
9. That the consumption of saturated fat and polymorphisms in the gene coding for Apolipoprotein E4 have independent and combined effects on serum cholesterol levels.

Further examples of hypothesised gene/exposure effects of current interest are available in the published literature[3, 4, 5] but a variety of additional hypotheses are likely to have come to light by the time nested studies are conducted.

### 2.1.5 Justification of study design

#### *2.1.5.1 Study design*

A variety of study designs can be used to investigate different aspects of the relationship between genotype, exposure and the risk of disease. A large scale population-based prospective study is the most appropriate setting for the comprehensive and reliable

quantification of the combined effects of genotype and exposure on a variety of outcomes. Such a project provides the infrastructure necessary to support a range of studies (particularly nested case-control studies) within the cohort and constitutes an important investment in the future of research in the area. The study proposed here has a number of advantages over other approaches, including advantages over a series of independent (i.e. non-nested) case-control studies. These features and their attributes are as follows:

1. Prospective cohort design;
- allowing investigation of a large number of different conditions/endpoints, including all-cause mortality;
- providing information on exposure prior to development of disease, thus avoiding recall bias and allowing accurate measurement of variables known to be affected by the disease process or by an individual's awareness of having a particular condition (e.g. weight, blood pressure, physical activity levels, diet);
- allowing the measurement of blood based molecular and proteomic factors (e.g. lipoproteins, hormones etc) using samples collected prior to the development of disease;
- providing prospectively collected blood samples, meaning that genetic information is available for all cases of disease in the cohort, regardless of the severity;
- allowing investigation of conditions which cannot generally be investigated retrospectively, (e.g. fatal conditions, dementia) and inclusion of all cases of disease where there is a high case-fatality rate (e.g. myocardial infarction);
- providing data allowing the broader consideration of both the risks and benefits associated with a specific genotype and/or exposure, through the inclusion of multiple endpoints;
- providing a straightforward source of comparable controls;
- minimising the assumptions made regarding the underlying relationship between genotype, exposure and outcome (in contrast to designs such as those employed in case-only studies, which assume independence between genotype and exposure);
- allowing the investigation of continuous outcomes;
- providing a framework for a variety of studies to be conducted within the cohort;
- providing a resource for the future, where investigation of outcomes and relationships unforeseen at the time of commencing the study is possible;
- providing a research resource which grows in value as time passes and more health events accrue.

2. Large-scale;
- yielding statistically reliable results;
- providing appropriate information on a range of important health outcomes;
- yielding accurate information on moderate effects, which are of clinical and public health relevance;
- providing an accurate and comprehensive quantification of combined genetic and environmental effects.

3. Population-based recruitment;
- allowing acquisition of information of direct relevance to health in the general population;
- providing a heterogeneous population, with a range of relevant exposures;
- providing direct information on rates of disease;
- allowing inclusion of large numbers of participants;

- encouraging a sense of public ownership and inclusivity.

4. Collaborative design;
- building infrastructure for high quality research in many different fields and institutions;
- allowing use of data by the broader biomedical community in the UK;
- making optimal use of existing expertise;
- providing a focus for training future geneticists, epidemiologists, genetic epidemiologists, statistical geneticists and other related disciplines.

5. Conducted through the UK National Health Service;
- allowing participation of the general population;
- making efficient use of existing resources;
- providing accurate information on health and other factors at the time of recruitment;
- providing accurate, unbiased computerised information on use of prescription medication;
- providing a centralised source of follow-up data;
- capable of providing accurate follow-up information on conditions which are important causes of morbidity but which do not necessarily result in hospitalisation or centralised registration for disease (e.g. diabetes);
- resulting in building of infrastructure for research;
- providing a direct interface between research and clinical practice.

Furthermore, as increasing numbers of endpoints are considered, the cohort study design becomes increasingly cost-efficient compared to stand-alone case-control studies. Estimates from the UK and the United States indicate that for stand-alone case-control studies with information comparable to the UK Biobank (i.e. questionnaire, interview, physical assessment and blood samples) the cost per case or control is in the region of at least £1000 (Day NE, Rothman N, personal communication). For incident diabetes mellitus, myocardial infarction, stroke, breast cancer, colon cancer and prostate cancer a total of 38,000 cases are expected in the UK Biobank after 10 years of follow-up (see section 2.3.11). Based on 3 controls per case, the total cost of comparable stand-alone case-control studies for these six important conditions is likely to be greater than that of the UK Biobank.

*2.1.5.2 Age range and sample size*

The cohort will consist of at least 500,000 men and women from the UK general population aged 45 to 69. The inclusion of individuals in this age range will allow the accrual of appropriate numbers of events within the required time frame (see section 2.3). The age range allows investigation of the common causes of morbidity and premature mortality and also allows ascertainment of events at an age where such cause-specific outcomes are generally well recorded, with less co-morbidity (and competing causes of mortality) than outcomes at older ages. The 25 year age range means that a variety of exposures can be examined (since exposures often vary by age) and that genotype/exposure effects can be examined stratified by age.

The decision to include at least 500,000 individuals is the result of the combined consideration of the number of events required for the reliable quantification of a number of important genotype/exposure/outcome relationships, (as outlined in section 2.3.11), as well as practical concerns regarding study design and cost. In particular, the inclusion of 500,000 participants will allow acquisition of sufficiently detailed exposure information while retaining feasibility within financial and organisational constraints.

It is also intended that the cohort will provide a national resource for biomedical research for several decades into the future. The chosen age range will provide a continuing stream of disease events for the next 20 or 30 years, acknowledging that events occurring after age 80 years will be of decreased informativeness due to co-morbidity and the generally observed decline of disease-exposure relative risks with age.

*2.1.5.3 Baseline measures and biological samples*
The baseline measures proposed for use in the UK Biobank are those which can be measured adequately and have been shown to be important in determining or ascertaining an individual's subsequent risk of disease (e.g. cigarette smoking, physical activity, blood pressure, weight etc), which are of current research interest (e.g. mobile phone use, sleeping behaviour etc), which must be taken into consideration when assessing genotype/exposure/outcome relationships or which are necessary to aspects of study methodology (e.g. information on family members).

Methods of collection, processing and storage have been devised to provide a robust and lasting source of DNA and plasma for future studies, taking cost into consideration. At the time of writing, the cost of immediate DNA extraction was estimated to be around £10 per participant. The working model for this protocol was assumed to be that DNA extraction would be deferred and done on a nested basis, when genotyping is required. This protocol therefore outlines the storage of genetic material as buffy coat frozen in liquid nitrogen. However, in recognition of the rapid changes taking place in this area, it is planned that this strategy be re-evaluated by experts in the area closer to the time when the more detailed methods and infrastructure for collection of biological samples are being put into place (probably in late 2002). By then, revised estimates of cost will be available, as will more extensive data from the EPIC study regarding the quality and quantity of DNA yielded from buffy coat held in liquid nitrogen for long term storage.

## 2.2 Background

### 2.2.1 Previous publications

Previous studies of the combined effects of genotype and environmental exposure on disease risk have not provided a consistent picture of the determinants of disease, for a number of reasons. Such studies have often been characterised by:

- small numbers of participants, leading to low statistical power and statistically unstable results. This has also led to apparently conflicting results between different studies.
- incomplete or inadequate measurement of exposure. In particular, many studies do not have appropriate controls, or do not have equivalent exposure information on both cases and controls.
- a lack of formal statistical testing for various effects, particularly statistical "interaction".
- the inability to provide a full picture of the combined effects of genotype and environmental exposure (see section 2.3.11 below). Many studies focus on very specific aspects of the relationship between genotype and environment (*e.g.* the "interaction ratio") and do not provide comprehensive information on the risk of disease associated with exposure in individuals with and without a particular genotype.
- use of a retrospective case-control design. These studies necessarily involve people who have survived a disease long enough to take part, do not provide information on rates of disease and use measures of exposure which are subject to recall bias.

### 2.2.2 Other large-scale studies with blood samples

*2.2.2.1 Country-based biological collections*

National biological sample collections are underway in Iceland and are planned in Estonia and Canada. These collections will provide relatively large quantities of genetic and other biological material and have considerable potential to improve the understanding of the relationship between genotype and the risk of disease. However, at this stage they do not involve the systematic prospective large scale collection of information regarding exposure. The collections are therefore limited in their ability to investigate combined genetic and environmental effects on the risk of disease. Furthermore, the population over 50 included in the Iceland study is relatively small.

*2.2.2.2 Cohort and other studies*

The study currently underway which is closest to the project proposed here in both size and design is the European Prospective Investigation into Cancer and Nutrition (EPIC), a prospective cohort study of men and women aged 40-74 at entry, based at 20 centres from 9 countries of the European Union plus Norway. Recruitment started in 1992, and blood samples, including DNA, have been stored on 370,000 of the cohort.

EPIC was designed to investigate the role of diet in the aetiology of the more common cancers. The endpoint follow-up should be adequate for mortality, by broad cause, and incident cancer cases. The coverage for non-cancer, nonfatal endpoints varies widely across the centres and for only a minority would there be nearly complete ascertainment of the types of endpoints which will be able to be investigated in the UK Biobank (see section 2.3.9). Exposure information was focussed on dietary variables, so some important lifestyle factors are inadequately covered, (e.g. physical activity), and most of the cohort did not undergo a clinical examination, so basic measures such as blood pressure and pulse rate are not available. The scope of the UK Biobank is therefore considerably wider than that of EPIC.

There are a number of studies underway in the UK which have aspects in common with the UK Biobank (see section 5). However, they generally have a narrower focus than the proposed study and all involve a smaller number of participants providing biological samples. The key feature of large scale comprehensive prospective information on genotype, exposure and outcome remains unique to the UK Biobank. Links are proposed to other studies, where appropriate (see section 5).

## 2.3 Plan of investigation

The UK Biobank involves the recruitment of 500,000 individuals aged 45-69 from the general population of the UK. These individuals will provide information on a range of exposures, will undergo an interview and physical examination and will provide a blood sample. Recruitment will be through general practices and will take place over a 5 year period. Study participants will be followed up for incident disease through population registries, hospitalisation records, general practice, a repeat questionnaire and other UK National Health Service sources for at least 10 years.

### 2.3.1 Structure of the proposed project

The project is collaborative in design, with a central “hub” and 5 or 6 “spokes”. The specific functions of the hub and spokes have not been finalised at the time of writing. However, it is envisaged that the hub will be responsible for the central administration of the project (including quality control), along with centralised storage of data and

biological material. The spokes would be responsible for liaison with general practices, recruitment of participants and the initial processing of biological material.

The study is designed as a national resource for researchers to investigate combined genotype and exposure effects on disease outcome and many outcomes will be able to be investigated in the future (see section 2.3.11.3). The study Overseeing Body and Scientific Management Group will develop policies on the use and analysis of the data from the project. Researchers will apply to the Scientific Management Group with proposals to investigate specific genotype/environment/outcome effects, which will be subject to peer review. They will be judged on a number of standard criteria, including scientific merit, technical feasibility, clinical/public health importance and resource implications. Ethical approval from the appropriate committee will also be required and all studies must conform to the relevant legal requirements.

### 2.3.2 Recruitment of participants

In the United Kingdom, virtually all members of the general population are registered with a general practitioner, through the National Health Service. All men and women aged 45 to 69 and registered at participating general practices will be sent an explanation of the study and an invitation to take part, signed by their general practitioner, accompanied by the study questionnaire and consent form. General practitioners will have been asked to check through the list of potential participants and to remove from the list anyone they consider inappropriate for inclusion in the study (e.g. those too ill to take part, institutionalised etc). Guidelines for exclusion will be given to general practitioners and, in the case of illness, it will be made clear that only those too ill to take part should be excluded and that that mere presence of a particular condition is not sufficient grounds for exclusion.

Those potential participants wishing to take part will be asked to complete the questionnaire and to telephone a freephone number or to return a reply-paid slip to arrange an appointment with a research nurse at the local study recruitment centre. There will be an interval of some weeks between receiving the questionnaire and consent form and attendance at the recruitment centre, allowing the potential participant time to consider their involvement in the study and to discuss aspects of the study and consent with family, friends and others. A freephone number will be available for any enquiries.

Participants will be asked to bring the completed questionnaire to their appointment with the research nurse. At this appointment the statement of consent will be discussed and those wishing to take part will provide written consent to do so. Selected parts of the self-administered questionnaire will be discussed and checked/verified by the research nurse. More detailed information regarding past medical history, current medication and identifying data (including NHS number) will also be sought. A number of physical measurements will then be taken, along with a blood sample. The discussion of consent, interview, physical assessment and blood sample is expected to take 30-45 minutes; the exact timing of the recruitment visit will be established more accurately in pilot studies (see section 2.3.13).

It is planned that recruitment clinics will serve a number of practices, in order to maximise efficiency. Such clinics are likely to be physically located outside of the general practice setting (e.g. in a mobile caravan unit) since most practices would not have sufficient space to house them. The clinic location and the conduct of recruitment will be tailored to the requirements of each practice. The study is designed to minimise

any disruption to the usual delivery of care at the participating general practices. Abnormalities detected at the initial physical assessment (e.g. hypertension, dysrhythmias etc) will be conveyed, in a standardised way, to the participant's general practitioner for further assessment (with consent of the participant, see sections 2.3.4.4 and 5.4.1).

### 2.3.3 Cohort composition

The cohort will be composed of people aged 45-69 from selected general practices who consent to take part. The cohort is not designed to be representative of the general population of the United Kingdom. Rather, it is planned that participants will come from a broad range of socioeconomic backgrounds and regions throughout the UK, with a wide range of exposure to factors of interest. Increased participation from certain groups will be achieved by selection at the level of the general practice, rather than selection within a practice (see below). Pilot studies will compare characteristics of participants and non-participants (see section 2.3.13). Spouse pairs will be encouraged to join the study and participants will be asked for the name, date of birth and address of their spouse, siblings and offspring. Record linkage techniques will be used to link family members within the cohort.

#### *2.3.3.1 Achievement of required sample size*

In the UK, the average general practice size is around 6,000-8,000 patients, although research oriented practices may be somewhat larger. The project will aim to recruit around 40 to 50% of the eligible population from each practice. This means that around 1-1.2 million individuals will need to be approached to provide the required sample of 500,000 participants; around 500-600 general practices would provide this number of individuals in the appropriate age range. To gauge the appropriateness of the target response rates, these will be assessed during the pilot studies and monitored throughout the recruitment period, and revised if necessary.

#### *2.3.3.2 Selection of general practices*

General practices will be considered eligible to take part in the study if they have an interest in research and use computerised prescription records. Over 90% of practices in the UK use such records. Practices will be selected so as to cover a broad range of regions and socio-economic conditions throughout the UK. In many cases, negotiations to take part in the study will be conducted through Primary Care Trusts, rather than through individual practices or practitioners.

#### *2.3.3.3 Age and sex*

Since it is likely that the effects of genotype and environmental exposure will need to be examined by age and sex, it is important that substantial numbers of participants are included in each 5 year age group, for both men and women. At this stage, it is not known what the age and sex composition of those aged 45-69 choosing to take part in the study from the practices is likely to be and the estimated numbers of events in the cohort (see section 2.3.11.3) has been based on recruitment of around 50,000 participants in each 5 year age and sex group. Demographic features of the age-group 45-69 in the UK (i.e. the post second world war "baby boom") mean that recruitment may be weighted towards younger age-groups. Pilot studies will be conducted prior to commencement of the main study to establish the likely participation profile and participation rates will be monitored throughout the course of the study.

*2.3.3.4 Ethnicity*

There will be no restriction on participation according to ethnicity, so the number of participants from various ethnic groups will reflect the composition of the general population at the study practices and the ethnic backgrounds of individuals choosing to join the study. This means that it is likely that around 90% of participants will be white.

The study aims to recruit sufficient numbers of participants from the larger ethnic groups (as defined by the 1991 census) to be able to ascertain the prevalences of certain exposures and genotypes within each group and to estimate to what extent disease and death rates in each group can be explained by the prevalences of these risk factors. It is planned to recruit at least 3,000 individuals within each group. This will be achieved by including an appropriate number of general practices serving the relevant ethnic groups. Study participation by ethnic group will be monitored over the course of the study and measures will be taken to ensure that these recruitment targets are met.

### 2.3.4 Baseline measures

Information gathered at baseline will cover a broad range of exposures. Table 1 lists the potential data to be collected in the questionnaire, interview and physical assessment. Emphasis will be placed on validated measures of each exposure. Detailed development of the study instruments (including questionnaire, interview, physical assessment and sample collection procedures) will take place in consultation with collaborating scientists and other experts (see section 3).

*2.3.4.1 Questionnaire*

For exposures which can be measured accurately by self-reporting, a self-administered questionnaire will be used. The participant will be mailed the questionnaire to complete at home and will be asked to bring the completed form to the appointment with the research nurse. Included in the questionnaire are personal characteristics which potentially influence the risk of disease and/or need to be taken into account when assessing the relationship between genotype, exposure and disease. They cover the broad areas of socio-economic status, demography, habits/lifestyle, diet, reproductive history, family history, past health, disability/impairment, psychological status and early life factors.

*2.3.4.2 Consent*

The consent form for participation and follow-up will be included with the questionnaire. Prior to the interview and examination, consent to join the study will be discussed and any queries that the potential participant has will be addressed, after which he or she will be asked to provide written consent to participate (see section 5.1). He or she will be reminded that participation in the study is purely voluntary and that he or she is free to withdraw from the study at any stage. It will also be made clear that any decision regarding participation will not affect his or her health care in any way. Potential participants will also be assured of the confidentiality of all personal data and that such data will be treated in full accordance with the Data Protection Act and will be used for medical research only. For potential participants with concerns about specific aspects of the study, it will be explained that participation entails taking part in all aspects of the study (i.e. completing the questionnaire, interview, physical assessment, donation of blood and follow-up) and that any person with misgivings about any aspect of the study should consider declining participation.

*2.3.4.3 Interview*

At the interview the research nurse will check through key aspects of the self-administered questionnaire to minimise missing data. The participant will then be asked about his or her medical and surgical history, and current medication use (along with dose and duration of use). Additional data on current prescribed medication will be obtained from the practice computerised prescription record at the time when participants are flagged at the practice (see section 2.3.9).

*2.3.4.4 Physical assessment*

Where applicable, physical measures will be taken according to standard criteria (e.g. measurement of blood pressure will be in accordance with standards set out by the British Hypertension Society[6]). The physical assessment will commence with the first measurement of blood pressure, followed by measurement of pulse rate, anthropometry (height, weight, waist and hip dimensions), and forced expiratory volume (FEV1). It will finish with the second measurement of blood pressure. Participants will receive feedback on the results of their physical assessment and any abnormalities found will be conveyed to their general practitioner (see section 5.4.1).

*2.3.4.5 Additional data on diet*

It is planned that having had their interview, examination and blood collection, all participants will be given a 7-day diet diary to complete at home. A reply paid envelope will be provided for its return to a centralised location; completed diaries will be stored and archived for future nested studies. These data on diet are additional to the food frequency questionnaire which will be included in the main self-administered questionnaire. While it is recognised that inclusion of the diaries costs around £1 million and such diaries are relatively costly to code when data on nutrients are required, they add substantially to the information obtained on diet, which is clearly an important exposure.

**2.3.5 Blood samples**

The general principles and procedures for the collection, processing and storage of blood samples from participants are outlined below. However, these will need to be considered in more detail as the study infrastructure is put in place and they remain open to potential modification, should the need arise.

*2.3.5.1 Requirements*

Since the main nested analyses of the material collected by the project will be conducted at least 10 years after recruitment, methods of collection, processing and storage of biological samples must ensure the long term stability of a range of analytes. Furthermore, sufficient quantities of material must be available for multiple studies and the material collected must be versatile enough to provide for a range of unforeseen research needs. The main requirement of the biological samples is to serve as a reliable long-term source of DNA for genotyping and material for biochemical, haematological and proteomic measures (see tables 2 and 3 for examples of the types of analytes which could potentially be measured using blood samples from the cohort).

The protocol for the collection, processing and storage of blood samples therefore has two main aims:

- to provide material for future studies. While DNA and frozen plasma are essential, the project is designed to maximise the number of analytes which will be able to be measured accurately in the future.

- to employ appropriate measures which allow for feasible large-scale recruitment, with optimal value for money.

Tables 2 and 3 list a number of important analytes, along with their known processing and storage requirements. These are included as examples of the types of analytes which may be measured within the project; they are not intended to indicate which analytes will necessarily be measured in the future, as these will be defined by the research needs at the time when nested studies are conducted. Table 2 indicates that a relatively large number of analytes remain stable if collected in disodium ethylene diamine tetraacetic acid (EDTA), stored at room temperature for up to 48 hours, then stored in liquid nitrogen (-196°C) in the longer term. Table 3 indicates that a number of haematological, haemostatic and biochemical analytes have additional requirements.

Chilling of blood samples immediately after collection improves the stability of many of the analytes listed in table 2 and is essential for the accurate measurement of others (e.g. creatinine, homocysteine, vitamin C etc). However, chilling at 4°C is not essential for all analytes (and is detrimental to the analysis of levels of factor VII) and it is therefore proposed that a proportion of the blood collected be stored and transported at 4°C, prior to processing and a proportion remain at room temperature. This could also be considered a safety measure, since having both chilled and unchilled samples improves the chances that various unforeseen analyses will be able to be performed successfully in the future. However, it is recognised that having chilled and unchilled samples, *versus* having only unchilled samples, costs around £1 million to £2 million overall and increases the complexity of the project.

In order to provide an extra long-term source of DNA for controls, a random sample of 10,000 members of the cohort will have an extra 5ml of blood taken for cryopreservation of peripheral blood lymphocytes in a manner which will allow subsequent immortalisation of these cells. Figure 1 illustrates the proposed method for collection, processing and storage of blood samples. It should be noted that although centrifuging samples at the point of collection is considered necessary for the measurement of the activity of certain clotting factors, it was not considered practical for the purposes of the project.

In addition to the known requirements, there are a number of uncertainties remaining regarding the optimal means of collection, processing and storage of blood samples for the subsequent measurement of emerging blood-based proteomic and metabonomic analytes. At present, the majority of known analytes can be measured in plasma with EDTA or citrate stored in liquid nitrogen. The added value of samples preserved with specific protease inhibitors (or other reagents and storage techniques) cannot be quantified at this stage, nor is it clear which would be the most appropriate to use. However, much work is currently underway or planned in this area and the project will remain open to the possibility of adapting the methods of blood collection, processing and storage, should the need arise.

*2.3.5.2 Collection*

In general, participants will not be fasting at the time when their blood is collected. However, the potential exists for those participants with morning appointments to be asked to attend the clinic in a fasting state, and the time since the last meal will be recorded for all participants.

A total of 50ml of blood will be collected from each participant; 40ml in EDTA (4 x 10ml), and 10ml in sodium citrate. The samples will be collected into vacutainers in the

following order: tube 1- EDTA, tube 2- citrate, tubes 3, 4 and 5- EDTA (figure 1). Tubes 1, 3 and 5 will be stored and transported to the processing centre at 4 °C and tubes 2 and 4 will be stored and transported at room temperature, prior to processing. If the full amount of blood cannot be collected then the tubes should be filled in the specified order until no more blood can be taken. Thus, if only 20ml of blood can be collected a chilled sample of blood with EDTA and a room temperature sample of blood with citrate would be available to the study. The exact time and date of collection will be recorded. Samples will be transported to the laboratory (at the "spoke") and processed within 48 hours of collection.

*2.3.5.3 Processing and long term storage*

*(i) Samples kept at room temperature*

The samples with EDTA and with citrate will be centrifuged at the processing centre and separated into aliquots of plasma, buffy coat and red blood cells (figure 1). All samples will be frozen at -80°C before being transferred to liquid nitrogen tanks for long term storage.

*(ii) Samples kept at 4°C*

A small amount of blood with EDTA from the chilled samples will be analysed immediately to provide a full blood count (including haemoglobin level, red cell count, white cell count and differential, packed cell volume, mean cell volume, mean cell haemoglobin, mean cell haemoglobin concentration and platelet count) on all participants. It is planned to use an automated coulter counter which samples directly from the vacuum tube by piercing the bung, allowing the remainder of the tube to be centrifuged and processed normally. All tubes will then be centrifuged. Some aliquots of the plasma will have 10% meta-phosphoric acid (MPA) added to them for measurement of vitamin C and the remainder will be aliquotted as EDTA-plasma only. Buffy coat and red blood cells will be aliquotted. All aliquots will be frozen at –80°C before being transferred to liquid nitrogen for long term storage.

*(iii) Cryopreservation of peripheral blood lymphocytes from a random sample of 10,000 members of the cohort*

Prior to blood collection, a random sample of around 2% of the cohort will have been identified to provide an extra 5ml blood sample for the cryopreservation of peripheral blood lymphocytes. This blood will be collected in acid citrate dextrose (ACD) and will be processed (including graded freezing) so as to allow the subsequent immortalisation of these cells, if necessary.

*2.3.5.4 Analysis*

The storage system used for the biological samples will ensure access for future studies nested within the overall cohort. When DNA is required for genotyping, frozen buffy coat aliquots will be thawed and DNA will be extracted. DNA which is surplus to the immediate requirements will be aliquotted and stored in liquid nitrogen. HbA1c levels will be measured in frozen buffy coat, which correlate extremely well with the levels found in fresh whole blood (Clark S, personal communication). Thawed plasma (with EDTA) will serve as the main source of material for biochemical and proteomic analytes and plasma (with citrate) will provide material for measurement of other analytes (such as certain haemostatic factors). The samples which were chilled prior to processing and storage in liquid nitrogen will provide material for the measurement of serum vitamin C (with EDTA and MPA), homocysteine and other analytes (e.g. table 3). Should the need arise for larger amounts of DNA from the random sample of study participants, their frozen peripheral blood lymphocytes will be revived and immortalised.

For all studies utilising data from biological samples, care will be taken to ensure that variables such as time between collection and freezing, years in storage, batch and measurement method are accounted for in comparisons of cases and controls.

*2.3.5.5 Data management, quality control and safety*
All samples will be labelled with the appropriate study identifier (using a barcode, where possible) and data management systems will be in place to record and track their collection, processing and storage. Data on the results of all analyses (including genotyping) will be stored linked to the study identifier (see section 2.3.10).

Strict quality control measures will be applied to ensure the quality of stored samples and biological data and to ensure comparability of samples between spokes and over time. Such quality control measures will be the responsibility of the hub. Each participant's biological material will be stored divided between two locations with independent storage systems, to minimise the risk of complete loss of biological material, should any failure of storage equipment occur.

### 2.3.6 Resurvey of a subset of participants to correct for regression dilution

In order to correct for regression dilution, independent sets of around 2,000 participants will undergo a repeat of all recruitment measures (questionnaire, interview and blood sampling) approximately every two years after recruitment.

### 2.3.7 Studies incorporating molecular, proteomic and metabonomic measures

An important and expanding area of research interest is that of investigating the relationship between genotype, levels of blood-based gene products (i.e. molecular, proteomic and metabonomic analytes etc) and outcome. Such substances have a wide range of likely applications, including serving as biomarkers of exposure, independent risk factors for disease, intermediate markers of disease phenotype, measures which can be used to classify disease status and as outcomes in their own right. Investigations focussing on blood-based measures are likely to yield important information regarding disease and mechanistic pathways as well as enhancing the ability of the project to quantify certain effects, through the accurate measurement of exposure and the minimisation of measurement error.

### 2.3.8 Pharmacogenetics

Due to the unique structure of the UK National Health Service, virtually all prescription medication taken by study participants will have been prescribed by their general practitioner and will be recorded in the computerised practice prescription database. Such data show very good agreement with self-reported information regarding treatment for various conditions, with kappa scores generally of 0.7 or higher[7]. These databases therefore provide a reliable, unbiased and accessible electronic source of data on prescription of medication to participants which can be updated regularly. At baseline, all medications currently prescribed to participants will be extracted electronically from the practice computerised record, along with their doses and duration of use. Research in this area is known to be of considerable interest to the commercial companies, particularly those involves in biotechnology and pharmaceutical development.

### 2.3.9 Follow-up procedures

The initial follow-up period is 10 years. It is acknowledged that considerable value would be obtained over the longer term, as the cohort ages and more events accrue.

*2.3.9.1 Follow-up through NHS central registers*

All study participants will be flagged through the Office of National Statistics and the study co-ordinators will therefore be notified routinely of deaths (with details of cause of death) and cancer registrations in the cohort. Additional follow-up through the Scottish Morbidity Register will be instituted for participants in Scotland.

*2.3.9.2 Follow-up through general practice, hospitalisation records and other NHS sources*

Study participants will also be followed-up regularly through their general practice for incident morbidity, including any episodes of hospitalisation. Due to the variety of electronic and paper systems used by different practices, the means of follow-up will have to be flexible and will need to be tailored to each practice. It is expected that the method of routine follow-up will range from manual retrieval by study staff for selected events in participants to a fully computerised enquiry of the practice system and a number of different methods of follow-up of participants will be investigated in pilot studies. In certain parts of the UK, it is also possible to link individuals to hospital records to ascertain episodes of hospitalisation, along with the reason for admission. It should be noted that these methods of follow-up are currently successfully employed by a number of different epidemiological studies being conducted through general practice (e.g. EPIC-Norfolk, studies carried out through the General Practice Research Framework etc) and that the study proposed here is not dependent on the development of new health information systems. However, over the course of the study the method of follow-up will be adapted to take advantage of advances in computerised recording, storage and retrieval of patient information; such advances are likely to improve the efficiency of follow-up. In particular, development and improvement of information systems within the National Health Service are underway or planned. The possibility of follow-up through direct linkage with hospital records will be explored.

*2.3.9.3 Follow-up and updating of exposure through questionnaires to participants*

The initial period of follow-up for the study is 10 years. During this time certain aspects of exposure (e.g. smoking, alcohol consumption, physical activity etc) are likely to change in a number of participants, generally resulting in dilution of measured effects of exposure on disease risk, due to misclassification. It is planned that all study participants will undergo some form of resurvey approximately 5 years after recruitment. This resurvey will include a brief postal questionnaire to ascertain self-reported incident morbidity and to update selected exposures. Participants' reports of incident morbidity will be validated and additional evidence of disease will be sought to allow classification of disease status according to standard criteria. It should be noted that funds for this resurvey are not being sought here; this section is included for information only.

*2.3.9.4 Ascertainment, validation and classification of diagnoses*

For most conditions, any general practice based or self-reported diagnosis will be taken as an initial indication only of the possible presence of disease. Further evidence of disease status will be sought (e.g. hospital records, pathology data etc) and diagnoses will be validated and classified according to standard criteria. For example, it is planned that general practice based or self-reported diagnosis of myocardial infarction will be confirmed using hospital and pathology records and will be classified according to World Health Organisation criteria.

The relative phenotypic heterogeneity of many of the conditions to be investigated by the study is acknowledged. It is planned that information will be obtained to allow the classification of diseases into appropriate pathological or severity related sub-categories.

For example, any suspected stroke will be confirmed using hospital records and information will be sought to allow classification into subgroups such as haemorrhagic, ischaemic (large vessel, small vessel) and undetermined etc. Staging information will be sought for cancers, where appropriate.

*2.3.9.5 Minimisation of loss to follow-up*

At the time of giving consent, all participants will have been asked for explicit written permission for follow-up and to contact them again in the future (see section 5.1). Over the course of the project, a number of participants will change address, potentially causing difficulties with follow-up. For those who move within the UK, follow-up for cancer and mortality through the NHS central register will not be affected. Loss to follow-up by other means will be minimised in a number of ways. First, participants will be asked to keep the study informed of any change of address. Second, the list of study participants (with NHS number) will be matched regularly against the Office of National Statistics lists of individuals registered at specific health regions (a process known as "list cleaning"). Any individuals found to have moved health regions will be re-contacted using updated address details. Finally, general practitioners will be asked to provide forwarding details of any "flagged" study participants who have changed practices.

**2.3.10 Data management**

The management of data on exposure, physical attributes, biological samples, outcomes and biochemical, proteomic and genotypic measurements will be a complex and large-scale undertaking. Extensive systems to ensure reliable, high quality and accessible data management will need to be developed for the project and systems will need to be standardised across the hub and spokes. It is envisaged that such systems will be developed and maintained in-house, to allow rapid responses to data quality issues and appropriate adaptation to the changing needs of the study. Experts on the development and maintenance of such systems will be involved from the earliest stages of planning of the study and appropriate funds for this and other information technology needs will be ensured. A more detailed informatics strategy is currently being developed.

*2.3.10.1 Recruitment*

A unique study identification number will be assigned at the point where potential participants are asked to join the study. All study forms and samples will be labelled with the study identifier and barcoded to facilitate handling of the data.

*2.3.10.2 Data entry*

All study forms will be checked, coded and scanned electronically to allow semi-automated data capture and storage of forms as computerised images. Data capture will be by means of specialised programmes (incorporating optical mark reading and intelligent character recognition etc) with operator verification and validation where appropriate.

*2.3.10.3 Confidential handling and storage of data*

All identifying details will be stored separately from the other data recorded for each participant. Participant data from the questionnaire, interview, samples and follow-up will be stored linked with the study identifier only. The file linking the study identifier with identifiable participant details will be kept under strict security, with access to authorised personnel only. Linkage of participant identifying information with their other study data will take place only when strictly necessary (e.g. to avoid sending out follow-up questionnaires to participants who have died, to identify and ask further questions of

participants with specific conditions etc) and guidelines for such linkage will be drawn up by the Scientific Management Committee prior to commencement of the study. Data released for analyses will not contain identifying information. All data will be handled in accordance with the Data Protection Act.

*2.3.10.4 Organisation*

Organisation of invitations to take part in the study will be the responsibility of the "spokes" and spokes will also be responsible for arranging study clinic appointments and documenting attendance at the clinic, return of study forms and samples. To maximise efficiency, the processing of study forms (including scanning and data entry) will be as centralised as possible. Storage of data from participants, including blood samples, will be centralised and will be the responsibility of the hub. Monitoring of recruitment rates and overall data quality will be the responsibility of the hub.

### 2.3.11 Statistical methods

*2.3.11.1 Analysis of combined effects of genotype and environmental exposure*

The main analyses to be conducted within the cohort will be defined in more detail by those researchers who are successful in applications to use the database, once recruitment and follow-up are sufficiently complete. At this stage, the main principles of analyses and power calculations are outlined. The main study design for assessment of the combined effects of environment and genotype consists of a series of case-control studies nested within the cohort. Options for the selection of controls include an individually matched design or a panel of controls selected at random from the cohort (probably weighted by age and sex). An important principle underlying the design of the study and its statistical methods is to minimise the assumptions made about the underlying nature of the relationship between genetic and environmental factors and the risk of disease.

In order to explore fully the combined effects of environment and genotype, the study will provide information about the risk of disease associated with exposures of interest, for individuals with and without a particular genotype, and will also give information regarding the effect of genotype on the risk of disease in people with and without particular exposures (tables 4a and 4b). This will allow a summary table of odds ratios to be produced (table 4c), where "A" represents the relative risk of disease in participants without the exposure or genotype of interest (usually used as the reference group), "B" and "C" represent the relative risk of disease according to genotype and exposure respectively, relative to group "A", and "D" represents the relative risk of disease in participants with both the genotype and exposure of interest. The interaction ratio relates to the effect of the exposure in individuals with and without the genotype of interest and is equal to (D/B)/(C/A). For the purpose of these power calculations, "statistical interaction" was considered to be departure from a multiplicative model, *i.e.* interaction ratio not equal to 1.0.

The need for information over and above that relating to formal statistical interaction comes from the observation that, in isolation, knowledge of this parameter and the "interaction ratio" are of limited value[8]. The relationship between the factor V "Leiden" mutation, oral contraceptive use and venous thromboembolism is a case in point; while no significant statistical interaction between genotype and exposure is apparent (with respect to a multiplicative model), their separate and combined effects are substantial and of direct clinical relevance, since exposed carriers of the mutation have a relative risk of disease of 35 (95% confidence interval 8-154) compared to non-exposed non-carriers[9].

For common diseases, common exposures and genotypes which result in a 20-30% increase in risk of disease can have a significant public health impact. The study is therefore designed to be able to detect relatively subtle but important effects of exposure and genotype on the risk of a range of common conditions. Common exposures and genotypes are considered to be those affecting around 20-80% of the population (see tables 5 and 6).

Although various classes of medication may be taken by over 20% of the study population (e.g. antihypertensive agents, non-steroidal anti-inflammatory drugs, exogenous hormones etc), pharmacogenetic studies tend to focus on the effects of more specific drug types (e.g. angiotensin converting enzyme inhibitors, statins, beta-blockers etc) or individual drugs. The prevalence of use of many individual drugs (e.g. omeprazole, paroxetine, insulin etc) is around or below 5% (unpublished data from the Million Women Study) and adverse drug effects (i.e. outcomes) are also relatively rare. Studies of combined drug/genotype effects must therefore be powered to address relatively low prevalences of exposure and low incidences of outcome. Bearing this in mind, the proposed study will have adequate power to address the risk of more common adverse drug effects in relation to commonly used medications and genotypes (see below).

For simplicity of calculation and presentation these power calculations focus on binary genotypes, exposures and outcomes. However, it should be noted that many situations will be more complex than this, in that these factors may be categorical or continuous. It is also appreciated that genetic risk covers a broad spectrum, with variable penetrances within and between genes. Furthermore, gene/gene interactions and exposure/exposure interactions within the relationships being examined may occur, and confounding may also be an issue. While binary genotype, exposure and outcome measures tend to result in relatively conservative estimates of power, increasing complexity also tends to increase the sample size required. It should be noted that the estimates of the numbers of events required do not take stratification or adjustments for environmental or genotypic covariates or exclusions of participants into account, nor do they account for missing data or measurement error. They therefore present the minimum number of events required for the detection of a particular effect under ideal circumstances; more events are likely to be needed since these issues are relevant to all studies of this nature.

In any analyses incorporating assessment of the relationship between genotype and outcome, careful account will be taken of participants' ethnicity and of population subgroup stratification. It is planned that ethnicity and population stratification will be classified in two main ways; according to questionnaire information on self-reported ethnic group (and country of origin) and according to characteristics of selected genetic markers in cases and controls. Where appropriate, analyses will be stratified by or matched on the resultant categories and subcategories of ethnicity/population subgroup. It should be noted that recent data from regional studies in the UK indicate that population stratification does not appear to have a large impact on the investigation of genotype/exposure/outcome relationships (Day NE, personal communication). Nevertheless, the issue of confounding by ethnicity/population subgroup is recognised as important in the reliable assessment of genotype/outcome relationships (and, consequently, to the investigation of combined genotype/exposure effects) and how this is dealt with within the UK Biobank will be the subject of extensive and ongoing

consultation with experts in the area, including statistical geneticists, genetic epidemiologists and others.

*2.3.11.2 Power and sample size calculations*

The study is designed with adequate power to estimate accurately the relative risks of disease associated with a particular exposure within the genetic subgroups of interest and the relative risks of disease associated with a particular genotype within each exposure group, the three relative risks of interest (B-D) and the interaction ratio (and whether or not this ratio is likely to differ significantly from 1.0). A 0.1% level of significance has been chosen due to the importance of avoiding a type I error and because of the number of statistical tests which are likely to be carried out. Calculation of the number of events required for accurate quantification of specific exposure/genotype/outcome relationships is not straightforward and is dependent on a number of different inputs and assumptions. These are:

1. Ratio of cases to controls (given for 1:4 and 1:10).
2. Required power (assumed to be 95%).
3. Level of significance (assumed to be two-sided, 0.1%).
4. Proportion of participants experiencing specific exposures (see table 5 for examples).
5. Proportion of participants with putative risk-conferring allele(s) (see table 6 for examples).
6. Required minimum/maximum detectable relative risk associated with exposure within genetic subgroups (see table 7).
7. Required minimum/maximum detectable relative risk associated with genotype within exposure subgroups (see table 7).
8. Required minimum/maximum detectable interaction ratio (see table 8 and figures 2a-d).
9. Assumptions regarding the underlying model of combined genotype/exposure effects (see table 8).

The calculations in table 7 were made using the EpiInfo statistical package (version 6) and those in table 8 and figures 2a-d used a programme developed by the National Cancer Institute (USA).[10,11] It is apparent from tables 7 and 8 and figures 2a-d that estimates of the number of events required for specific minimum relative risks and interaction ratios are sensitive to the prevalences of the exposures and genotypes of interest. Furthermore, the number of events required is affected by the choice of model for the combined effect of the genotype and exposure. Since the likely nature of the underlying relationship between genotype, exposure and risk of disease is not known at this stage, it is important that power calculations account for a range of possibilities.

*2.3.11.3 Expected numbers of events in the cohort*

Table 9 presents the expected number of events in a cohort of 500,000 people aged 45-69 after 5 and 10 years of follow-up if disease and death rates are similar to those in other UK cohort studies. Where possible, these estimates take account of the "healthy cohort effect" and exclusion of individuals with prevalent disease at baseline. For example, rates of incident myocardial infarction in the cohort are assumed to be 50% of those in the general population and rates of lung cancer are assumed to be 30% of those in the general population. The assumptions on which these numbers are based are given in Appendix 1. It should be noted that this table provides approximate numbers of events only.

*2.3.11.4 Discussion of power*

Assuming at least 20 to 80% of participants are exposed to the factor of interest and 20 to 80% have the genotype of interest, for outcomes with around 5,000 events the study should be in a position to detect relative risks of disease associated with a particular genotype and/or exposure of at least 1.5 and an accompanying interaction ratio of around 1.4 (tables 7 and 8, figures 2a-d). For outcomes with around 1,000-2,000 events, the study should be in a position to detect relative risks of disease associated with a particular genotype and/or exposure of at least 1.8 to 2.0 and an accompanying interaction ratio of 1.7 to 2.0. Greater power would exist for more common exposures and genotypes and lesser power would be present for less common exposures and genotypes.

This means that for genotypes and exposures present in 20-80% of the cohort and outcomes with at least 5,000 events, the study should be in a position to detect the effects of modifiers of risk in genotypic and exposure subgroups with a population attributable fraction of at least 9%. The corresponding figure for events occurring in 1,000-2,000 participants is around 10-13%. Broadly speaking, this means that the study would be able to detect the effect of genotypes and exposures which are responsible for several hundred deaths in the UK annually. For example, a 10% attributable fraction for the risk of prostate cancer means that the factor under examination would account for around 500-700 deaths from prostate cancer per year in the UK.

Data contained in table 9 indicate that, for a cohort aged 45-69 at recruitment, 500,000 participants can be expected to yield a number of important outcomes with around at least 5,000 incident cases or deaths (e.g. diabetes mellitus, ischaemic heart disease mortality, myocardial infarction, stroke, colorectal cancer, breast cancer) and many others with at least 1,000 events (e.g. rheumatoid arthritis, Parkinson's disease, hip fracture, prostate cancer, bladder cancer, etc), over 10 years of follow-up.

*2.3.11.5 Investigation of combined genotype and exposure effects on continuous outcomes/phenotypes*

Considerable potential exists within the UK Biobank to examine the combined effects of genotype and exposure on a range of continuous outcomes, based on data gathered at recruitment. Such outcomes could include questionnaire-based measures (e.g. anxiety/depression scores, age at menopause, reported alcohol consumption etc), physical assessment findings (e.g. body mass index, waist/hip ratio, blood pressure etc) and the large number of measures based on analysis of blood samples (e.g. white cell count, HbA1c, cholesterol level, fibrinogen levels, hormone levels etc).

Because of the nature of these types of outcomes, relatively highly powered studies are possible using cross-sectional data from thousands and tens of thousands of participants. Furthermore, the number of participants involved in investigations of genotype/exposure effects could be varied according to the measures and outcomes of interest; the precise nature of the studies to be conducted will be defined by researchers who are successful in bidding to use data from the UK Biobank. The number of participants required to investigate reliably effects on continuous outcomes is considerably smaller than the number required to generate the dichotomous outcomes of interest for nested case-control studies. It can therefore be assumed that the cohort of 500,000 will provide sufficient data for these genotypes, exposures and continuous outcomes to be investigated at very high levels of power and very stringent levels of significance, for relatively rare exposures and genotypes. For a detailed example of the power provided for evaluation of continuous phenotypes by a sample of 40,000

individuals within the cohort, readers are directed to the accompanying illustrative protocol for a more intensively phenotyped sub-cohort.

These studies will provide important information on a number of continuous traits directly implicated in the development of disease. At the same time, they will allow analyses to be conducted relatively early in the recruitment period, providing timely information relevant to the interpretation of disease-based endpoints and maintaining scientific interest and prospects for innovation for collaborating scientists.

### 2.3.12 Training of study staff

A large number of staff will be involved in the project and the long term nature of the study as well as the need for uniform, comparable and standarised acquisition, processing and storage of data underlie the importance of careful and ongoing training of staff. In particular, research nurses will receive extensive training to ensure high quality data collection according to standard criteria. Assessment of inter- and intra-observer variability of measurements will be made during pilot studies and, where necessary, at other times during recruitment. Any unacceptable levels of variability will be addressed by measures such as revision of protocols and/or additional training. Systems of quality control will be put in place at the commencement of the study and will be the responsibility of the hub.

### 2.3.13 Pilot studies

Prior to commencement of the main study, pilot studies will be conducted to establish the feasibility of recruitment, to optimise study methodology and to provide information regarding recruitment and other factors, to aid in the planning and conduct of the main study.

The main pilot study will involve commencement of the main study at a small number of practices with close monitoring of response rates, participation profile, data quality and technical issues. This pilot study will also be used to test aspects of the information technology requirements of the study as well as methods of follow-up. Comparisons will be made between study participants and non-participants, using data such as postcode (to derive indices of deprivation) and age. Characteristics of the participants will also be compared with published population-based data on exposures such as smoking and alcohol consumption (e.g. from the Health Survey for England). Calls to the study helpline will be carefully documented and participants and staff involved with the study will have the opportunity to comment on its conduct. Where indicated, the protocol for the main study will be revised in the light of the pilot study findings.

### 2.3.14 Other studies within the framework of the cohort

The UK Biobank will provide an ideal and cost-effective framework for smaller, more detailed studies of combined genotype/exposure effects. The Protocol Development Committee for the UK Biobank strongly supports the conduct of such studies within the cohort, in particular it recommended study of intensive phenotyping outlined in the accompanying document. However, this and other studies should be considered separately from the main proposal for their scientific merit and funding; the overall proposal for the UK Biobank deals with the data to be collected on all 500,000 participants and subsequent nested studies and is not dependent for its success on additional studies within the cohort.

Bearing this in mind, smaller studies within the UK Biobank will be able to focus on the examination of exposures, phenotypes and outcomes which would not be feasible in the

whole cohort, but which are nonetheless important in the understanding of genotype/exposure effects and their underlying mechanisms of action. By focussing on appropriate continuous measures of exposure and outcome and on minimising measurement error, accurate and reliable quantification of certain genotype/exposure/outcome measures can be achieved by studies involving thousands or tens of thousands (rather than hundreds of thousands) of participants, within the larger study cohort. The potential also exists for studies focussing on participants identified as having specific conditions (e.g. investigations of factors determining disease progression or survival), exposures or familial relationships (e.g. studies focussing on siblings and/or spouses). It is also likely that novel sub-study designs will emerge as the study progresses.

A general principle of sub-studies taking place within the cohort is that they must fit in with the overall study design and must not compromise acquisition of the universal dataset or conduct of the study in general. Any proposals for studies within the cohort must meet with the approval of the study Scientific Management Group, Overseeing Body and appropriate ethical committees.

### 2.3.15 Logistics and timetable

This document outlines the scientific case for the first 5 years of support for the UK Biobank, which is expected to cover the period from April 2002 to April 2007. Figure 3 shows the projected timetable for the study from January 2002 to the end of 2008, to allow a broader view of the project.

Contingent on favourable peer-review, funding and ethical committee decisions, it is planned that the process of identifying the hub and spokes will commence around April 2002 (following on from calls for expressions of interest, which will take place in December 2001). Applicants will be required to submit proposals outlining their suitability for the role, including the availability of suitable academic, informatic and physical resources. In parallel with this process, the Overseeing Body and Scientific Management Group will be set up and the chief executive officer of the project will be recruited.

The setting up of the infrastructure for the project (including making ready premises, recruitment of staff, developing informatics etc) will begin around July 2002. Initial pilot studies, which will represent a smaller scale commencement of the overall study, are planned for the last quarter of 2002. The commencement of recruitment into pilot studies will be accompanied by the appropriate data entry and blood sample processing and storage. Flagging through general practice and the Office of National Statistics will be tested during pilot studies and will commence in full soon after the start of recruitment into the main study. Identification and recruitment of appropriate general practices by the spokes will commence in the last quarter of 2002, in preparation for the main study recruitment, which is scheduled to begin in April 2003.

Follow-up through the Office of National Statistics for mortality and cancer registrations will be ongoing. Biennial follow-up through general practice will commence in early 2005 and the questionnaire to update exposure and to ascertain incident self-reported morbidity will be sent out to participants approximately 5 years after recruitment, commencing in early 2008. Recruitment of participants is expected to be completed in April 2008. If the median time of recruitment is taken to be late 2005, then a median of 10 years of follow-up on the cohort will have been achieved by late 2015.

## 3. FURTHER DEVELOPMENT OF STUDY INSTRUMENTS AND INFRASTRUCTURE

This protocol outlines the general principles and scientific case for the UK Biobank. While a number of details of the study will require further development (outlined below), many aspects of the study can be considered to be settled at the time of writing. These include the fact that the proposed project will involve the setting up of a prospective population-based study involving 500,000 men and women aged 45-69 in the UK. Recruitment will be through general practice and will involve a self-administered questionnaire, interview, physical assessment and blood collection. Baseline measures will cover those broad areas outlined in table 1 and follow-up will be through the National Health Service Central Register and general practice, as well as other methods, such as linkage with hospital admissions.

Bearing in mind the general principles and features of the study, further development of the study instruments and infrastructure will be required before the study can be conducted. It is planned that collaborating scientists involved with the study "spokes" and appropriate external researchers will form a pool of experts and detailed development of the study instruments (including questionnaire, interview, physical assessment and biological sample collection procedures) and infrastructure will occur in consultation with these experts. In particular, experts will include specialists on questionnaire design and study methodology, database development, information technology, epidemiologists, geneticists, genetic epidemiologists, statistical geneticists, clinical chemists etc, to ensure that optimal and practicable methods are employed. It is also planned that collaborators and researchers with expertise on specific diseases be encouraged to set up disease related networks to develop and guide appropriate disease oriented research.

In refining the main potential baseline and outcome measures and translating them into appropriate study instruments, both scientific and pragmatic concerns will be taken into consideration. Due to the large scale of the resources which would be invested in the UK Biobank, emphasis will be placed on the use of validated measures of exposure and methods of sample collection and storage which have been shown to be reliable in the long term (see section 2.1.5.3). Development of the instruments will also take into account the needs of study participants and it will be ensured that the questionnaire, interview, physical examination and blood collection procedures will not be overly demanding and will allow widespread participation, at the same time as satisfying the main scientific goals of the study.

## 4. EXPECTED OUTCOMES OF THE STUDY AND CONTRIBUTION TO KNOWLEDGE

As the largest and most comprehensive prospective study with biological samples in the world, the UK Biobank is expected to contribute substantially to international knowledge regarding the combined effects of genotype and exposure on the risk of disease. Its design means that the study will provide a structure and resources for future research, and will enable researchers to address current and unforeseen scientific questions.

### 4.1 Health and wealth implications

The availability to the research community and to industry of large scale information on genotype, exposure and outcome will provide unprecedented opportunities for discovery and innovation in genetics, epidemiology, biotechnology and pharmacology. The UK Biobank will provide the resources necessary for the translation of aspects of basic scientific knowledge from the laboratory to the broader health context and from a

theoretical understanding of genotype and exposure effects on disease aetiology to reliable empirical evidence regarding their observed combined effects on human health.

For the UK, BioBank represents a substantial, broad and accessible investment in post-genome research. By spreading the task of recruitment over a number of research centres, and by making data available to the general scientific community, the project will serve as a means to develop national expertise and infrastructure in genetic and molecular epidemiology. It will provide a focus and resource for high quality research over the coming decades. Collaboration with industry will allow research and investment to translate findings from the project into products and innovations of direct benefit to the general population.

## 5. ETHICAL CONSIDERATIONS

The conduct of the UK Biobank will conform to the relevant ethical and legal guidelines regarding consent, confidentiality and the use of human tissue and biological samples. The project will undergo extensive ethical review and will be required to meet the approval of the appropriate UK Multi-Centre Research Ethics Committee. The project will be conducted in accordance with relevant aspects of the Human Rights Act 1998, the General Medical Council's Guidance on Confidentiality and the Council of Europe's Recommendation on the Protection of Medical Data. It will also follow the guidance outlined in the Medical Research Council's documents on Personal Information in Medical Research and on Collections of Human Tissue and Biological Samples for Use in Medical Research. Ethical aspects of the project are the subject of ongoing consultation and will be dealt with in greater detail elsewhere, in the future; this document therefore provides an outline only of the main ethical considerations.

### 5.1 Informed consent

Prior to the request for consent, potential participants will have been informed of the following in writing and will have had the opportunity to discuss them with a research nurse and via the freephone helpline:

- the purpose and nature of the study;
- the study methods and what participation would involve;
- that participation is purely voluntary and that they are free to withdraw from the study at any time;
- that any decision they make about participation will not affect their future health care in any way;
- that their data will be kept strictly confidential in accordance with the Data Protection Act, used for medical research only and will never be used in a way which would identify them personally;
- that they will not receive any individual feedback about their blood results or any other aspect of their data, apart from the immediate result of their physical examination (see below);
- that the study will be important for future research and that many of the tests and analyses which will be conducted in the future cannot be specified at present;
- that their blood samples will be stored indefinitely;
- that their future health will be tracked through a number of different National Health Service sources;
- that they are likely to be asked to provide more information for the study in the future;
- that the research has been approved by the appropriate Multi-centre Research Ethics Committee and that all research carried out within the study will conform to strict ethical guidelines;

- that there will be a committee to monitor the conduct of the study which is independent of the study investigators and will ensure that the public interest is served by the study.

Potential participants will have a number of weeks between receiving information about the study and being asked to take part, to consider this information and their involvement in the study. If they indicate that they would like to take part they are asked to provide written consent for follow-up through NHS registers, their general practice and other medical records, for permission to use their data and blood samples for various analyses and specified and unspecified biochemical and genetic tests and for permission to contact them again at a later date. They will also be asked for permission to convey the results of their initial physical assessment to their general practitioner. It will be explained that their data will be available in anonymised form to approved researchers who submit successful applications to the Scientific Management Committee and that such researchers could include those from industry (see section 5.4.2).

## 5.2 Confidentiality

All data relating to the proposed study will be stored and used in an anonymised linked format and will be handled in strict accordance with the Data Protection Act throughout (see section 2.3.10). Data released for analyses will not contain any identifying information and publications from the study will not identify any individuals taking part. Great care will be taken to ensure the confidentiality of all study data and the risk to participants of breach of confidentiality is considered very low. Data from the project will not be accessible to the insurance industry or any other similar body.

## 5.3 Risk/benefit assessment

### 5.3.1 Risks

Since the proposed study is observational in nature, the study is essentially of low risk to participants. The main possible adverse effects relate to the consequences of completing the study questionnaire, interview and examination, having the blood sample taken and the risk of a breach of confidentiality (see above).

The questionnaire and interview will cover standard demographic, lifestyle, diet and health related questions (see table 1) and great care will be taken to avoid questions which are offensive or overly intrusive. Many questions will be similar to those used in existing studies. During the pilot studies the acceptability of the study methods will be carefully monitored (see section 2.3.13) and any problematic questions will be modified or removed before starting the main study. Potential participants will be given a freephone number to call with any comments or questions and calls to this number will be monitored to check for any ongoing problems. Interviews and physical examinations will be conducted by trained research nurses according to a standard protocol. The interview will be similar to most medical interviews regarding past health and, given the medical context of the study, should not be considered inappropriate or offensive. The examination does not contain any invasive or painful procedures.

Venepuncture is a safe and routine medical procedure which will be performed by experienced research nurses under standard medical conditions. However, the procedure may be painful and can result in bruising, although care will be taken to minimise these effects. Potential participants will be informed that a blood sample is required in order to take part in the study so that those unwilling to undergo venepuncture can decline participation. Venepuncture is considered essential to the

study, since a large number of important factors can only be measured reliably in blood and the amount of DNA which can be obtained from other sources (e.g. buccal swabs) is unlikely to be sufficient for the purposes of the study.

### 5.3.2 Benefits

In the longer term, the main potential benefit to people taking part in the study is that its findings may result in improvements in health and health care in the future. This is likely to benefit participants and other members of the general public, as well as future generations. In the short term, every participant will have their blood pressure and pulse measured and those with abnormalities will be referred to their general practitioner. It is likely that a number of people will be found to have undiagnosed hypertension and/or dysrhythmias and such people are likely to benefit from diagnosis and treatment. However, it will be made clear to potential participants that the study is not a "health check" and that any concerns that they have about their health should be raised with the appropriate health professional.

## 5.4 Additional ethical considerations

### 5.4.1 Feedback to participants

Participants will receive feedback only on measures taken at the physical assessment and any abnormal finding at this assessment will also be communicated to their general practitioner (with their consent). It will be made clear that they will not receive routinely any individual information relating to their blood samples (including blood biochemistry and genetic findings) or any other aspect of their study data. However, individuals will have the legal right to access their personal data, if required. Individuals concerned about their risk of particular conditions will be encouraged to consult their general practitioner or other health professional.

### 5.4.2 Commercial involvement

Major potential benefits of the project include the development of new treatments and the more precise targeting of existing therapy. Involvement of the pharmaceutical and biotechnology industry in the project is therefore essential to maximise delivery of potential health benefits. Any use of material from the study by commercial organisations will be subject to the approval of the Scientific Management Committee and the Overseeing Body and must conform to the relevant ethical and legal requirements. Potential participants will be informed of the likely involvement of commercial bodies, along with the reasons underlying their involvement.

## 6. LINKS WITH OTHER PROJECTS

A number of projects are currently underway in the UK which potentially overlap with the proposed project in terms of recruitment, scientific goals and other aspects (e.g. EPIC, The Million Women Study, The Whitehall Study, ALSPAC, ProtecT). There will be close liaison between this study and other studies working on similar areas to take advantage of potential links and to minimise problems associated with overlap.

## 7. DISSEMINATION OF FINDINGS

Dissemination of findings will be mainly by means of reports published in peer-reviewed journals. The Department of Health and the appropriate regulatory body will be informed at the earliest possible stage of any relevant findings. Overall relevant findings will be conveyed to study participants via regular newsletters and, where appropriate, web-based media will be used to make findings available to the public.

## Table 1. Potential baseline measures

| exposure category | variable/exposure | details required |
|---|---|---|
| **1. self-administered questionnaire** | | |
| socioeconomic/ demographic | age/date of birth | |
| | address/postcode | |
| | marital status (with details of spouse) | |
| | education | |
| | occupation/social position | |
| | ethnicity | birthplace of parents and grandparents |
| habits/lifestyle | smoking history | including onset, offset, duration, quantity smoked |
| | environmental tobacco smoke | including partner's smoking habits |
| | alcohol consumption | including quantity, pattern, type, past history |
| | physical activity | validated measure will be used |
| | sleep | hours per night, daytime nap |
| | mobile phone use | duration, frequency |
| | handedness | |
| diet | general diet/dietary changes | food frequency questionnaire, diet diary |
| | supplements | |
| | hot drinks | |
| reproductive history | number of children (plus dates of birth) | birthweight, gestational age, breastfeeding |
| | menarche/menopause or andrarche | |
| | use of exogenous hormones | |
| | hysterectomy/oophorectomy | |
| | vasectomy, tubal ligation | |
| family | family history of specific conditions | |
| | siblings | names, dates of birth, addresses (with postcode), twins |
| past health | medical and surgical history | past history of specific conditions |
| | previous head injury | |
| | screening behaviour | |
| | diagnostic/therapeutic radiation | |
| | over the counter medication use | |
| disability/impairment | activities of daily living | |
| | eyesight | |
| | hearing | |
| | Rose angina questionnaire | |
| | self-reported general health | |
| psychological status | measure of psychological profile | Hospital Anxiety and Depression Scale |
| early life | birth weight | |
| | weight history | |
| **2. nurse interview** | | |
| discussion of consent | | |
| NHS number | | |
| medical history | illnesses | |
| | medication history | details of current medications from practice database |
| | surgical history | |
| **3. examination** | | |
| anthropometric | height | |
| | weight | |
| | waist/hip | |
| biophysical | blood pressure | 2 measurements, 5 minutes apart, Omron machine |
| | pulse | |
| | forced expiratory volume | |
| **4. blood sample** | | |

## Table 2. Examples of analytes for which there is evidence of good stability if kept at room temperature for up to 48 hours, then long term liquid nitrogen

| source of specimen | analyte | comment |
|---|---|---|
| buffy coat | DNA | DNA extraction conducted on thawed buffy coat |
| | HbA1c | |
| plasma (EDTA) | total cholesterol | |
| | high density lipoprotein cholesterol | |
| | low density lipoprotein cholesterol | |
| | apolipoprotein A1 | |
| | apolipoprotein B | |
| | triglycerides | |
| | steroid hormones | |
| | peptide hormones | |
| | antibodies | |
| | fibrinogen | |
| | fibrin D-dimer | |
| | plasminogen | |
| | von Willebrands factor | |
| | antithrombin III | |
| | TPA antigen | |
| | C reactive protein | |
| | lipoprotein [a] | |
| | serum amyloid A | |
| | albumin | |
| | $\alpha$-carotene | better if kept in dark, effect of long term freezing uncertain |
| | $\beta$-carotene | |
| | cryptoxanthin | |
| | lutein | |
| | lycopene | |
| | retinol | |
| | $\alpha$-tocopherol | |
| | $\gamma$-tocopheral | |
| | free fatty acids | |
| | vitamin B12 | |
| | folate | some loss of stability with prolonged storage at room temperature |
| | cotinine | |

## Table 3. Examples of analytes with special requirements

| source of specimen | analyte | immediate requirements | | long term requirements | comment |
|---|---|---|---|---|---|
| | | processing | storage/transport | | |
| whole blood | haemoglobin<br>white cell count<br>white cell differentials<br>red cell count<br>MCV (PCV, MCH, MCHC)<br>platelet count | EDTA | | | must be tested fresh |
| plasma | homocysteine | EDTA | chilled | liquid nitrogen | |
| | creatinine<br>creatinine kinase<br>GGT<br>ALT<br>AST | EDTA<br><br><br>lithium heparin | chilled | liquid nitrogen | <br><br><br><br>very unstable |
| | vitamin C | EDTA | chilled | MPA<br>liquid nitrogen | ?long term stability |
| | factor VII<br>factor VIII<br>factor V<br>factor IX<br>PAI 1 antigen | citrate | not chilled<br>early centrifuge<br>early centrifuge | liquid nitrogen | |

**Table 4a. Schema for examination of the risk of disease associated with exposure in those with and without a particular genotype, where letters a to d refer to the number of cases and controls in each genotype/exposure category.**

**Genotype 1**

| | control | case |
|---|---|---|
| unexposed | $a_1$ | $b_1$ |
| exposed | $c_1$ | $d_1$ |

OR (C in table 4c)= $\frac{a_1 d_1}{b_1 c_1}$

**Genotype 2**

| | control | case |
|---|---|---|
| unexposed | $a_2$ | $b_2$ |
| exposed | $c_2$ | $d_2$ |

OR (D/B in table 4c)= $\frac{a_2 d_2}{b_2 c_2}$

**Table 4b. Schema for examination of the risk of disease associated with genotype in those with and without a particular exposure, where letters a to d refer to the number of cases and controls in each genotype/exposure category.**

**unexposed**

| | control | case |
|---|---|---|
| genotype 1 | $a_1$ | $b_1$ |
| genotype 2 | $a_2$ | $b_2$ |

OR (B in table 4c)= $\frac{a_1 b_2}{a_2 b_1}$

**exposed**

| | control | case |
|---|---|---|
| genotype 1 | $c_1$ | $d_1$ |
| genotype 2 | $c_2$ | $d_2$ |

OR (D/C in table 4c)= $\frac{c_1 d_2}{c_2 d_1}$

**Table 4c. Schema for examination of the combined effects of exposure and genotype, where letters A to D refer to the odds ratios of disease in each genotype/exposure category.**

| | genotype 1 | genotype 2 |
|---|---|---|
| unexposed | A | B |
| exposed | C | D |

interaction ratio= $\frac{D/B}{C/A}$

## Table 5. Proportion of UK women aged 50-64 with various exposures[12]

| exposure variable | frequency |
|---|---|
| smoking | |
| never | 50% |
| past | 26% |
| current | 18% |
| alcohol consumption | |
| none | 23% |
| <7 units per week | 54% |
| ≥7 units per week | 23% |
| strenuous exercise | |
| never/rarely | 44% |
| ≤ once per week | 30% |
| more than once per week | 26% |
| body mass index | |
| <20kg/$m^2$ | 5% |
| 20-25kg/$m^2$ | 47% |
| 26-29kg/$m^2$ | 34% |
| ≥30kg/$m^2$ | 15% |
| current treatment for: | |
| hypertension | 14% |
| high blood cholesterol | 6% |
| heart disease | 3% |
| diabetes | 2% |
| asthma | 7% |
| osteoarthritis | 8% |
| family history of breast cancer (mother or sister) | |
| yes | 10% |
| no | 90% |
| **risk factors applying to women only:** | |
| nulliparous | 11% |
| parous | 89% |
| tubal ligation | |
| no | 80% |
| yes | 20% |
| use of oral contraceptives | |
| never | 46% |
| ever | 54% |
| use of hormone replacement therapy | |
| never | 51% |
| past | 16% |
| current | 33% |

**Table 6. Proportions of controls with various "rare" or risk conferring alleles, from published studies**

| gene/polymorphism | range of proportion of controls with "rare" or putative risk conferring allele | reference |
|---|---|---|
| Apolipoprotein ε4 (Arg112-Cys) | 21% | 13 |
| BRCA1 | | |
| Pro871Leu | 32% | 14 |
| Gln356Arg | 7% | 14 |
| CAPN10 SNP43 | 85% | 15 |
| COMT Val158Met | 35% - 54% | 14 |
| CYP1A1 | | |
| Ile462Val | 4%-9% | 14 |
| 3'UTR T6235C | 10% - 12% | 14 |
| Thr461Asp | 4% | 14 |
| CYP2D6 poor metaboliser | 4% - 12% | 14 |
| CYP2E1 C carrier | 7% - 9% | 14 |
| CYP17 promoter T→C | 38% - 42% | 14 |
| CYP19 $(TTTA)_n$ | 0.5% - 2% | 14 |
| EDH17B2 Ser312Gly | 47% | 14 |
| ER XbasI RFLP | 68% | 14 |
| ER CCC325CCG | 13% - 21% | 14 |
| Factor V Arg506Gly (Leiden) | 5% | 9 |
| GSTM1 deletion | 38% - 51% | 14 |
| GSTP1 Ile105Val | 28% - 29% | 14 |
| GSTT1 deletion | 21% - 27% | 14 |
| HSP70-hom N1/N2 | 41% | 14 |
| HSP70-2 P1/P2 | 4% | 14 |
| INS VNTR | 30% | 16 |
| NAT1 A1088T | 17% - 47% | 14 |
| NAT2 slow acetylator | 49% - 62% | 14 |
| PR PROGINS | 13% - 18% | 14 |
| TD2*PPARG Pro12Ala | 80% | 17 |
| TP53 | | |
| intron 3 16-bp insertion | 12% - 16% | 14 |
| intron 6 G→A | 10% - 56% | 14 |
| Arg72Pro | 26% - 35% | 14 |
| TNF-α 308T→A | 17% | 14 |

**Table 7. Number of events required for examination of the separate effects of exposure and genotype on the risk of disease, according to various prevalences of exposure and genotype and minimum detectable relative risks (assuming 95% power, and 0.1% significance).**

| proportion of exposed controls | proportion of participants with genotype of interest | minimum detectable relative risk of disease associated with exposure, in those with and without genotype of interest* | minimum detectable relative risk of disease associated with genotype, in those with and without exposure of interest** | number of events required | |
|---|---|---|---|---|---|
| | | | | 4 controls per case | 10 controls per case |
| 50% | 50% | 1.3 | 1.3 | 3600 | 3200 |
| | | 1.5 | 1.5 | 1500 | 1300 |
| | | 1.8 | 1.8 | 750 | 700 |
| | | 2.0 | 2.0 | 550 | 500 |
| 20% | 20% | 1.3 | 1.3 | 12700 | 11200 |
| | | 1.4 | 1.4 | 7600 | 6650 |
| | | 1.5 | 1.5 | 5150 | 4500 |
| | | 1.8 | 1.8 | 2350 | 2000 |
| | | 2.0 | 2.0 | 1650 | 1400 |
| 10% | 10% | 1.6 | 1.6 | 12600 | 11000 |
| | | 1.8 | 1.8 | 7700 | 6650 |
| | | 2.0 | 2.0 | 5300 | 4600 |

*see table 4a, odds ratios C and D/B
**see table 4b, odds ratios B and D/C

**Table 8. Number of events required for detection of various interaction ratios, according to various prevalences of exposure and genotype (assuming 95% power, 0.1% significance and 10 controls per case).**

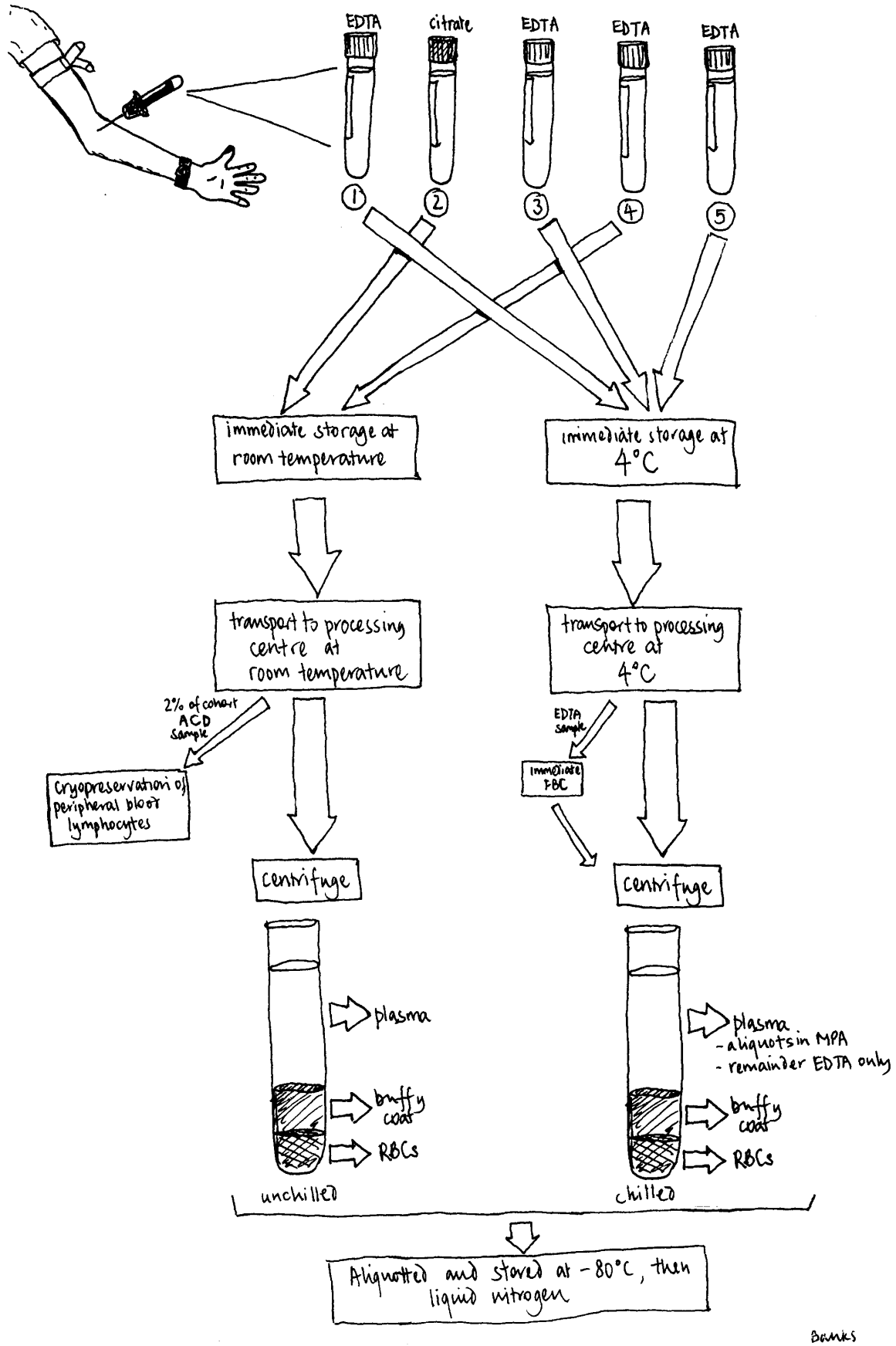
| | exposure | | genotype | | minimum detectable | number of events required[1] | |
|---|---|---|---|---|---|---|---|
| | prevalence | relative risk | prevalence | relative risk | interaction ratio | 4 controls per case | 10 controls per case |
| variation in prevalence of exposure and genotype | 50% | 1.5 | 50% | 1.5 | 1.5 | 3200 | 2900 |
| | 20% | | 20% | | | 4300 | 3500 |
| | 10% | | 10% | | | 11300 | 8900 |
| | 5% | | 20% | | | 12500 | 9950 |
| variation in relative risk and interaction ratio | 20% | 1.5 | 20% | 1.5 | 2.0 | 1400 | 1100 |
| | | | | | 1.6 | 3150 | 2600 |
| | | | | | 1.4 | 6350 | 5200 |
| | | | | | 1.3 | 10500 | 8800 |
| | | 1.4 | | 1.4 | 1.5 | 4500 | 3700 |
| | | | | | 1.4 | 6600 | 5500 |
| | | 1.3 | | 1.3 | 1.5 | 4700 | 3900 |
| | | | | | 1.4 | 6950 | 5800 |
| | | | | | 1.3 | 12000 | 9800 |
| | 10% | 1.5 | 10% | 1.5 | 1.7 | 6300 | 4900 |
| | | | | | 1.6 | 8200 | 6450 |
| | 5% | 1.5 | 20% | 1.5 | 1.8 | 5600 | 4400 |
| | | | | | 2.0 | 3900 | 3000 |

1. Model assumes that risk spread through population i.e. that relative risks for exposure apply to null genotype and that relative risks for genotype apply to non-exposed group

**Table 9. Estimated numbers of events among a cohort of 500,000 people aged 45-69**

| condition | total number of cases expected | |
|---|---|---|
| | after 5 years of follow-up | after 10 years of follow-up |
| **ASCERTAINABLE FROM ROUTINE DATA** | | |
| Ischaemic heart disease (deaths) | 2200 | 4300 |
| Cerebrovascular disease (deaths) | 540 | 1100 |
| Chronic obstructive airways disease (deaths) | 290 | 590 |
| Breast cancer (incidence) | 3010 | 6270 |
| Colorectal cancer (incidence) | 2140 | 5410 |
| Prostate cancer (incidence) | 1090 | 3290 |
| Lung cancer (incidence) | 980 | 2540 |
| Non-Hodgkin's lymphoma (incidence) | 610 | 1410 |
| Bladder cancer (incidence) | 430 | 1130 |
| Ovarian cancer (incidence) | 510 | 1120 |
| Stomach cancer (incidence) | 325 | 855 |
| **ASCERTAINABLE THROUGH GENERAL PRACTICE AND OTHER FOLLOW-UP** | | |
| Diabetes mellitus (incidence) | 5800 | 11500 |
| Myocardial infarction (incidence) | 4100 | 8250 |
| Stroke (incidence) | 2250 | 4500 |
| Dementia (incidence) | 1000 | 2000 |
| Rheumatoid arthritis (incidence) | 850 | 1700 |
| Parkinson’s disease (incidence) | 650 | 1300 |
| Hip fracture (incidence) | 600 | 1200 |

**Figure 1. Procedure for collection, processing and storage of blood**




EDTA disodium ethylene diamine tetraacetic acid
ACD acid citrate dextrose
MPA meta phosphoric acid
RBCs red blood cells

**Figure 2a. Power for 50% prevalence of exposure and genotype, odds ratio of 1.5 for exposure and genotype and 0.1% significance**

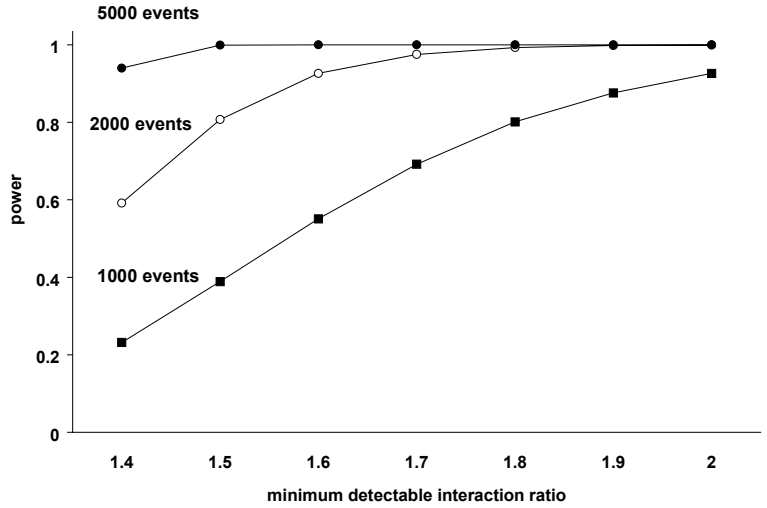



**Figure 2b. Power for 20% prevalence of exposure and genotype, odds ratio of 1.5 for exposure and genotype and 0.1% significance**

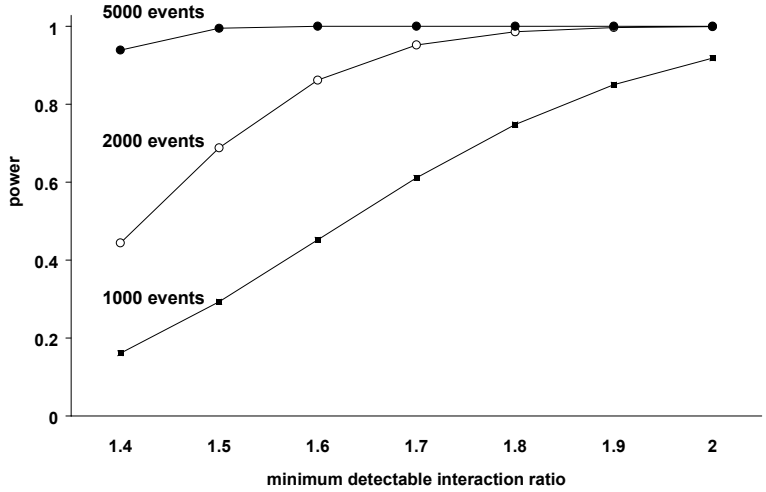



**Figure 2c. Power for 10% prevalence of exposure and genotype, odds ratio of 1.5 for exposure and genotype and 0.1% significance**

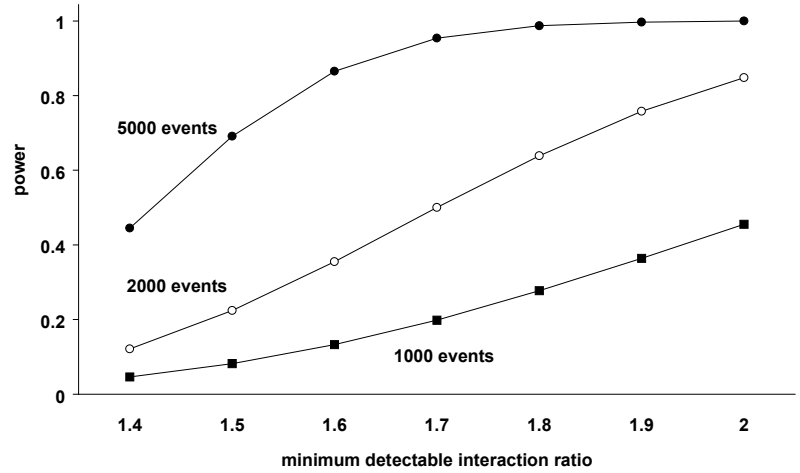
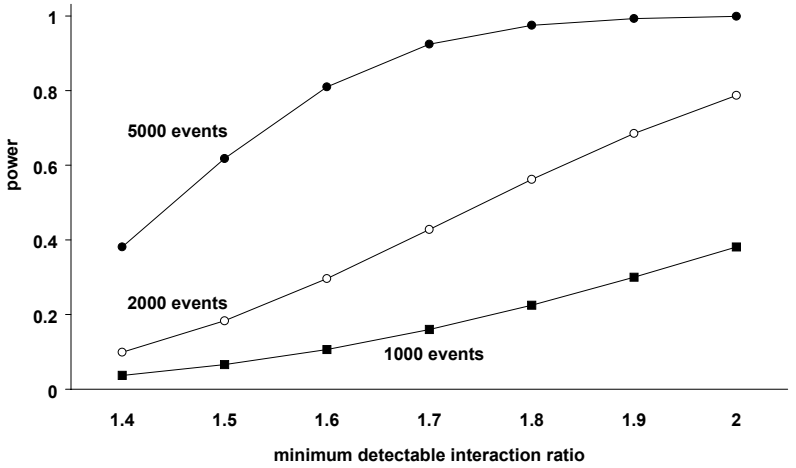
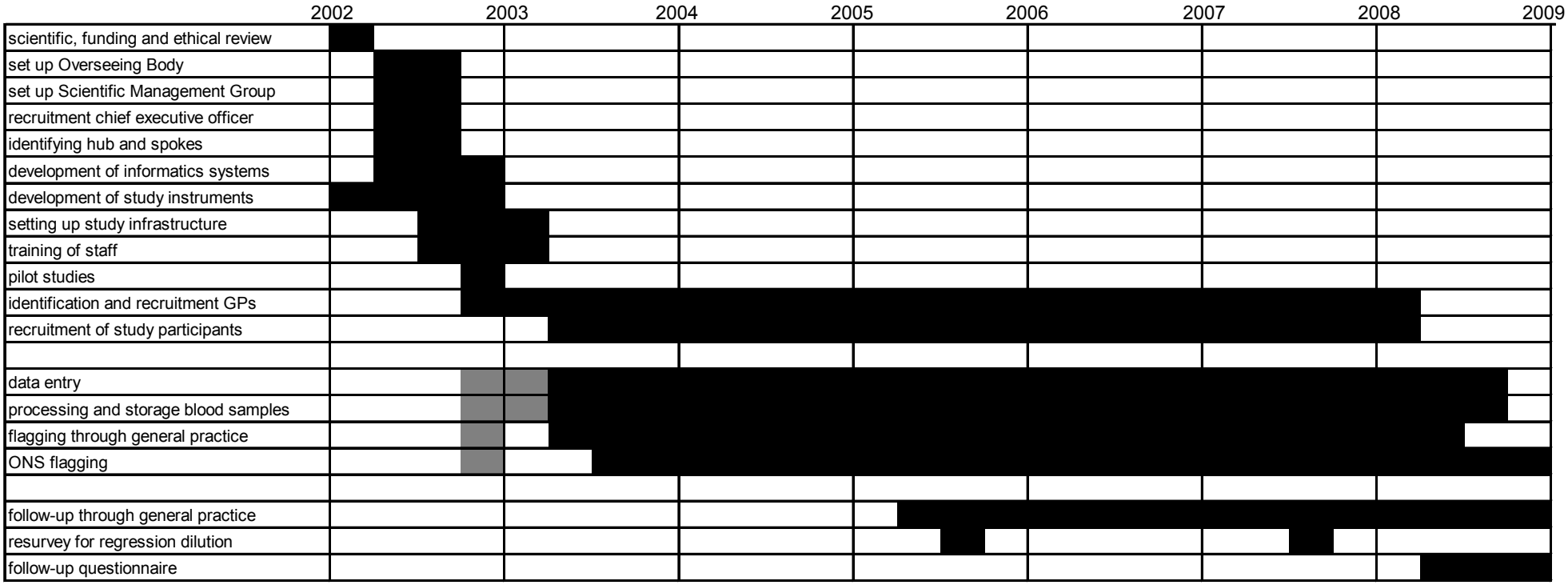



**Figure 2d. Power for 5% prevalence of exposure, 20% prevalence of genotype, odds ratio of 1.5 for exposure and genotype and 0.1% significance**



## Figure 3. Projected timetable for the UK Biobank, 2002-2009




GP general practitioner
ONS office of national statistics

testing during pilot studies

**APPENDIX 1**

The data displayed in table 9 are based on the following assumptions:

1. The age and sex distributions within the cohort were assumed to be even over the 25 year recruitment range (i.e. 50,000 men and 50,000 women in each 5-year age group).
2. The numbers of incident cancers expected were based on age, sex and cause-specific cancer registrations in England and Wales in 1994.
3. The numbers of deaths expected were based on age, sex and cause-specific mortality rates in England and Wales in 1995.
4. The numbers of incident hip fractures over a 10 year period were based on age and sex-specific incidence figures for 1983.[18]
5. For colorectal cancer, breast cancer, prostate cancer, ovarian cancer and non-Hodgkin's lymphoma, rates in the cohort were considered likely to be similar to those in the general population (based on unpublished data from EPIC and The Million Women Study).
6. Rates of bladder cancer, stomach cancer, pancreatic cancer and hip fracture in the cohort were considered to be 50% of those in the general population (based on unpublished data from EPIC and The Million Women Study).
7. Lung cancer rates in the cohort were considered to be 30% of those in the general population (based on unpublished data from EPIC and The Million Women Study).
8. For deaths from ischaemic heart disease and cerebrovascular disease, rates in the cohort were considered to be 50% of those in the general population and deaths rates due to chronic obstructive airways disease were considered to be 30% of those in the general population. Numbers of deaths were further reduced by 40% to allow for exclusion of participants reporting ischaemic heart disease, cerebrovascular disease and cancer at the time of recruitment (based on unpublished data from EPIC and the Million Women Study).
9. For Rheumatoid arthritis, data are based on the incidence of inflammatory arthritis reported to the Norfolk Arthritis Register and satisfying the American Rheumatism Association criteria for rheumatoid arthritis[19,20] (and Silman, A.J. personal communication).
10. The incidence of diabetes was based on data from the Health Survey for England, assuming that the incidence of disease is equal to the difference between prevalences in each successive age/sex group.
11. Data regarding the incidence of stroke (cerebral infarction, primary intracerebral haemorrhage and subarachnoid haemorrhage) were obtained from the Oxford Community Stroke Project[21] and those for myocardial infarction were taken from the Oxford Myocardial Infarction Incidence Study[22] (includes non-fatal and fatal definite myocardial infarction, fatal possible myocardial infarction and unclassifiable coronary deaths- MONICA definition 1). Rates in the cohort were taken to be those for men and women aged 45-64 and were considered to be 50% of those given for the general population of Oxfordshire.
12. Data on the incidence of dementia are based on information from a recent meta-anlaysis[23] and reduced by 75% to account for underascertainment in general practice.
13. The figures given for the incidence of Parkinson's disease are based on the incidence of parkinsonism in Carlisle, UK in men and women aged 40-69 from 1955-1961.[24]

## REFERENCES


1. Bell J. The new genetics in clinical practice. *Br Med J* 1998; **316**: 618-620.

2. Fears R, Roberts D, Poste G. Rational or rationed medicine? The promise of genetics for improved clinical practice. *Br Med J* 2000; **320**: 933-935.

3. Mucci LA, Wedren S, Tamim RM, Trichopoulos D, Adami HO. The role of gene-environment interaction in the aetiology of human cancer: examples from cancers of the large bowel, lung and breast. *J Intern Med* 2001; **249**: 477-493.

4. Rothman N, Wacholder S, Caporaso NE, Garcia-Closas M, Buetow K, Fraumeni Jr JF. The use of common genetic polymorphisms to enhance the epidemiologic study of environmental carcinogens. *Biochim Biophys Acta* 2001; **1471**: C1-C10

5. Ellsworth DL, Sholinsky P, Jaquish C, Fabsitz RR, Maniolo TA. Coronary heart disease. At the interface of molecular genetics and preventive medicine. *Am J Prev Med* 1999; **16**: 122-133.

6. O'Brien et al. Blood pressure measurement: recommendations of the British Hypertension Society. London: BMJ Publishing Group, 1997;

7. Banks E, Beral V, Cameron R, et al. Agreement between general practice prescription data and self-reported use of hormone replacement therapy and treatment for various illnesses. *J Epidemiol Biostatistics* 2001; **6**: 357-363.

8. Clayton D, McKeigue PM. Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet* 2001; **358**: 1356-1360.

9. Vandenbroucke JP, Koster T, Briet PH, Bertina RM, Rosendaal FR. Increased risk of venous thrombosis in oral-contraceptive users who are carriers of factor V Leiden mutation. *Lancet* 1994; **344**: 1453-1457.

10. Lubin J, Gail M. On power and sample size for studying features of the relative odds of disease. *Am J Epidemiol* 1990; **131**: 552-566.

11. Garcia Webb P, Lubin J. Power and sample size calculations in case-control studies of gene-environmental interactions: Comments on different approaches. Am J Epidemiol 1999;149:689-693. *Am J Epidemiol* 1999; **149**: 689-693.

12. The Million Women Study Collaborative Group. The Million Women Study: design and characteristics of the study population. *Breast Cancer Res* 1999; **1**: 73-80.

13. Lahoz C, Schaefer EJ, Cupples LA, et al. Apolipoprotein E genotype and cardiovascular disease in the Framingham Heart Study. *Atherosclerosis* 2001; **154**: 529-537.

14. Dunning AM, Healey CS, Pharoah PDP, Teare MD, Ponder BA, Easton D. A systematic review of genetic polymorphisms and breast cancer risk. *Cancer Epidemiol Biomarkers Prev* 1999; **8**: 843-854.

15. Horikawa Y, Oda N, Cox NJ, et al. Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nature Genetics* 2000; **26**: 163-175.

16. Huxtable SJ, Saker SJ, Haddad L, et al. Analysis of parent-offspring trios provides evidence for linkage and asssociation between the insulin gene and type 2 diabetes mediated exclusively through paternally transmitted class III variable number tandem repeat alleles. *Diabetes* 2000; **49**: 126-130.

17. Altshuler D, Hirschhorn JN, Klannemark M, et al. The common PPARgamma Pro12Ala polymorphism I is associated with decrease risk of type 2 diabetes. *Nature Genetics* 2000; **26**: 76-80.

18. Boyce WJ, Vessey MP. Rising incidence of fracture of the proximal femur. *Lancet* 1985; **i**: 150-151.

19. Symmons DPM, Barrett EM, Bankhead C, Chakravarty K, Scott DGI, Silman AJ. The incidence of rheumatoid arthritis in the United Kingdom: results from the Norfolk Arthritis Register. *British Journal of Rheumatology* 1994; **33**: 735-739.

20. Arnett FC, Edworthy SM, Liang MH, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis and Rheumatism* 1988; **31**: 315-324.

21. Malmgren R, Bamford J, Warlow C, Sandercock P, Slattery J. Projecting the number of patients with first ever strokes and patients newly handicapped by stroke in England and Wales. *Br Med J* 1989; **298**: 656-660.

22. Volmink JA, Newton JN, Hicks NR, et al. Coronary event and case fatality rates in an English population: results from the Oxford myocardial infarction incidence study. *Heart* 1998;**80**:40-44.

23. Jorm AF, Jolley D. The incidence of dementia: a meta-analysis. *Neurology* 1998; **51**: 728-733.

24. Brewis M, Poskanzer DC, Rolland C, Miller H. Neurological disease in an English City. *Acta Neurol Scand* 1966; **42 (Suppl 24)**: 1-89.