Informatics in Primary Care (2003) **11**: 129–35

Refereed papers

How could primary care meet the informatics needs of UK Biobank? A Scottish proposal

Frank M Sullivan PhD FRCP FRCGP
Professor of General Practice, Health Informatics Centre, University of Dundee, Dundee, UK

Jill P Pell MD FFPHM MRCGP
Consultant in Public Health, Greater Glasgow NHS Board and University of Glasgow and Director of the Scottish RCC, UK Biobank, Glasgow, UK

Mary Sweetland BSc MBA RSS
Deputy Director, Information and Statistics Division, NHS Scotland Common Services Agency, Edinburgh, UK

Andrew D Morris MD FRCP
Professor of Diabetic Medicine, Health Informatics Centre, University of Dundee, Dundee, UK

ABSTRACT

UK Biobank is an ambitious post-genomic project involving the recruitment and follow-up of 500 000 volunteers aged 45 to 69 years. Many primary care teams will be involved in the study directly or indirectly. The programme of research will use at least five data sources: paper-based questionnaires, blood samples, genotype information derived from the bloods, clinical/prescribing data from the medical records, and data on deaths. We describe three of the key challenges to primary care informatics posed by this project: patient recruitment, confidentiality, and data management. We then describe solutions proposed in Scotland, based on existing technologies. Some of these may be applicable elsewhere in the other Regional Collaborating Centres and other large-scale collaborative projects which rely on primary care informatics.

Keywords: confidentiality, data management, recruitment, research, retention

Introduction

The post-genomic challenge is coming to consulting rooms and clinics throughout the United Kingdom (UK).[1] We believe that the project offers four potentially vital prizes:

- insights into disease mechanisms in the families and individuals we look after
- a basic resource for science-based therapeutic innovation
- better targeting of preventive and therapeutic interventions to those most likely to benefit
- building up research capacity in primary care.[2–4]

Whether primary care engages with the process or not, the post-genomic revolution will roll on. Clinicians will be faced with patients who have, or perceive themselves to have, a health problem related to the new knowledge being generated by this revolution.

UK Biobank is a prospective cohort study recruiting around 500 000 volunteers aged 45–69 years.[5] Patients may be identified from general practitioner (GP) records or health board lists and invited to attend for baseline interview and examination. The project will be overseen by a co-ordinating centre in Manchester. This will be established as a charitable company limited by guarantee, to be owned jointly by the Wellcome Trust and the Medical Research Council

(MRC). There will be six regional collaborating centres (RCCs), which are responsible for recruitment around the UK:

- Scottish Consortium
- All Wales Consortium
- Central England Consortium
- Fosse Way Consortium
- London Biobank Consortium
- North West–Wessex Consortium.

The first visit will involve collection of questionnaire data on health and lifestyle, clinical examination, blood sampling for genetic and biochemical analyses and follow-up. Subsequent health outcomes will be ascertained via postal questionnaires and National Health Service (NHS) medical records, in particular those held at the general practice level. Therefore, close liaison with participating general practices will be essential, not only at the outset but throughout the follow-up phase. This paper describes three of the principal challenges to primary care informatics posed by UK Biobank: patient recruitment, confidentiality and data management. We then describe solutions proposed in Scotland, many of which might be applicable elsewhere in the other RCCs.

Patient recruitment and retention challenges

No decision has yet been taken about whether study subjects will be recruited by direct communication or through their GPs. If recruitment is via a network of GPs, then data will need to be acquired about the patients from records held on practice computer systems. If so, some consideration should be given as to which practices are recruited.[6] This might include eligibility criteria such as:

- a commitment to quality data recording and audit thereof
- an initial data quality audit
- a commitment to ensure that the electronic records for patients recruited are brought fully up to date
- the type of GP systems used must be able to provide the necessary data.

One of the critical success factors of the Biobank project will be the completeness and accuracy of the patient data collected.[7] Good primary care information technology (IT) solutions will be needed throughout the UK as patients move from place to place, no matter how initial recruitment occurs.

Confidentiality challenges

The UK Data Protection Act of 1998 has led to major difficulties for everyone engaged in post-genomic and epidemiological research using patient data.[8,9] Many believe that methods of working within the legislation can only be definitively clarified by the courts; however, few investigators will choose that route.[10,11] Debate as to what research activities may be legally undertaken has occurred in research ethics committees, scientific grant awarding bodies, the General Medical Council, the British Medical Association and the Royal Colleges, but a unified opinion has not emerged.[12,13] Unsurprisingly, Caldicott Guardians and other controllers of subject-identifiable patient data have been reluctant to release permission for access to data for fear of acting unlawfully.[14] Failure to achieve a satisfactory resolution would deprive researchers of access to the strongest and least biased methodologies, and patients of the benefits to patient care from projects like UK Biobank.[15] However, because Biobank participants are volunteers, many of the barriers can be overcome provided the consent process includes access to the appropriate data.

Data management challenges

Biobank will use at least four data sources in the study:

- paper-based questionnaires
- blood samples
  genotype information derived from the bloods
- clinical/prescribing data from the medical records of the cohort
- death data.

The validity, completeness, metadata, structure, format and accessibility of data in primary care varies depending on source.[16] Not all historical data may be entered, especially as the cohort will be recruited from an older population (45–69). There are also problems arising from the need to follow up patients in the longer term.

Proposed patient recruitment and retention solutions

At each participating practice, the master list of patients may be interrogated to produce a list of eligible subjects aged between 45 and 69 years. In Scotland,

all general practices use the Community Health Number (CHNo) to maintain practice lists. Individual GPs would then be asked to inspect the list of patients generated for their own practice. They would remove from this list the name of any individual they feel it would be inappropriate to contact on behalf of the UK Biobank project. Reasons for removal could include inability to provide informed consent and terminal illness. Exclusion may introduce socioeconomic bias and result in the study population being unrepresentative, therefore strict criteria will be used to minimise bias. Project staff would only get patient details once an affirmative response is received to the GP letter.

Invitations to participate

Letters could be prepared on the headed notepaper of individual practices. The letter of invitation to participate in UK Biobank would be signed by the GP to facilitate recruitment. We know from qualitative pilot work with patients in the target group that almost all patients approached are likely to agree to participate.[17] Monthly and quarterly review of recruitment rates would be conducted to identify and address problem areas.

Confidentiality solutions

Consent

At the screening visit each individual would be invited to consent specifically to:

- participate in UK Biobank, including initial interview, physical assessment, blood sampling and additional data collection on diet
- be contacted again in the future and invited to take part in further postal assessment or additional studies
- be agreeable to the manual inspection and recording of historical routine health data and information held in general practice and hospital records for the purpose of follow-up
- the automated record linkage of information held in computerised medical records in general practices and hospitals throughout NHS Scotland.

Security and encryption

The clinical information would be stored on a Structured Query Language (SQL) database. The first level of security on clinical data is imposed by the NHSnet. A log-in screen would be presented and all communications encrypted. Access would be determined by username and password. All attempts to log on to the Scottish RCC server and every action subsequently taken would be logged, producing audit trails. All clinical data will be anonymised with all patient-identifiable information removed before being passed for analyses, according to the UK Biobank protocol. This system has already been employed by the Health Informatics Centre and MEMO (Medicines Monitoring Unit) at the University of Dundee, and was consulted as part of the Confidentiality, Security and Advisory Group's report on patient confidentiality in NHS Scotland. Thus, every patient, GP, general practice and hospital clinic is allocated a unique, randomly generated identifier, intended for the purposes of anonymous research. Researchers are then restricted to 'views' of data that use these anonymous identifiers and omit any identifying fields. Thus, all demographic information is removed except sex, age (in months) and social class (Carstairs Index). For genetic research, only the nurse performing the fieldwork has access to the CHNo. This nurse does not have access to any other patient-specific data other than those collected during direct patient contact. In addition, though the system administrator (through necessity) has access to all data on the server, the identity of individual genetic markers would not be available to them.

Informed consent

Regardless of any of the above, the patient is ultimately in control. The SQL system fully supports informed consent though the use of consent granting and denial forms that are scanned directly into the central database using optical character recognition. Should consent be denied, all records pertaining to that patient, with the exception of that patient's unique identifiers, are deleted. If consent is subsequently denied, any new data for that patient would be automatically deleted.

A key part of the Scottish RCC proposal is that electronic record linkage and fieldwork means that once enrolled, longitudinal follow-up is automated to allow tracking of all phenotypic end-points (for example, disease progression, all drug use, all primary care consultations, hospital admissions, investigations and death). It is also possible through the CHNo to track patients who migrate to other regions of Scotland to reduce losses to follow-up. We anticipate the development of methods to ensure Scottish procedures are compatible with the other five regions.

Data handling solutions

Baseline UK Biobank: fieldwork data collection

Following written informed consent, all clinical data collected at the screening visit would be checked, coded and scanned electronically to allow semi-automated data capture and storage of forms as computerised images. This would be co-ordinated by the hub.

UK Biobank Scottish RCC: automated data follow-up

Automated follow-up of electronic records of individuals is a novel feature of the Scottish component of UK Biobank. This would create an extremely efficient infrastructure for future research using UK Biobank data collected in Scotland. This would be enabled by the use of the CHNo and our established track record in record linkage of general practice, hospital, pharmacy and the Registrar General datasets. The NHS Scotland IM&T Strategy that is currently being implemented in every trust in Scotland advocates the administration of electronic health records through centralised record linkage techniques, using the CHNo as the unique patient identifier.[18] This is the Scottish Care Information (SCI) initiative. This strategy would clearly support the automated follow-up of all individuals enrolled into UK Biobank in Scotland.

Role of CHNo in facilitating automated follow-up

Every person who is registered with a general practice in Scotland is allocated a unique identifying number (CHNo). This is a ten-digit integer, the first six digits being the date of birth. Therefore every resident who is registered with a GP appears in the continuously updated computerised record, the Community Health Index (CHI). This file contains data on address, postcode, GP, deceased persons and date of death. Thus, the demographic breakdown of Scotland, death and patient migration can be easily analysed. The CHNo is used as the patient identifier in *all* primary healthcare activities in Scotland. In some areas (Dumfries and Galloway and Tayside), it is already used for *all* healthcare events in primary and secondary care.

Commitment of NHS Scotland to use the CHNo

NHS Scotland is committed to the principle that all people can easily register for general NHS Scotland services, specifically with a GP and dental practice, and change aspects of that registration with minimal impact on continuity of care, including a mechanism to facilitate the transfer of a GP patient summary from their old practice to the new. In addition there is a commitment that *core identification data* are maintained, including CHNo, in such a way that all contributors to care can use the data when and wherever required. Currently, the existing Community Health Index is being developed towards a system to be known as the SCI Index. A major part of this is the synchronisation of all principal master patient indices (MPI) used by primary care, secondary care and national systems. This is achieved by ensuring all MPIs carry the CHNo and that these are electronically generated from the main CHI as the source index. Obtaining electronic access to the main CHI through a number of approved access methods now available does this. The SCI initiative has used these techniques to implement a product called 'SCI Store' that is a central repository of all patient data. Clinical information is collected from legacy systems and stored in SCI Store and coded according to SNOMED-CT. It is currently installed in various stages in 11 out of 15 health boards within Scotland.

Linkage of historical data

Where the CHNo number is not available (such as for historical data), probability matching can be used, whereby a number of patient-identifiable fields are linked to calculate an overall probability of a match being due to chance, and a predetermined cut-off is then applied. The record linkage unit within the Information Statistics Division (ISD) of NHS Scotland has extensive experience of using probability matching to link research databases to routine admission and death data. This method has been used frequently within Scotland to provide follow-up information on admissions and deaths in research studies.

Proposed automated linkage to be developed in Scotland

The UK Biobank Scottish RCC proposal involves the establishment of one database server running SQL Server 2000 and one web server. This 'consortium' server would manage all data collected from the four Scottish centres that are partners in this Scottish RCC.

Flagging of UK Biobank participants

Following written informed consent by UK Biobank participants, an electronic 'flag' could be attached to the CHNo of that individual in the electronic systems that hold relevant healthcare information. These would be:

- the primary care legacy system (GPASS in the majority of practices)
- all SCI Stores in NHS Scotland
- the Scottish Morbidity Record (SMR) returns made to ISD.

We propose that information from these systems would be extracted on a nightly basis, using an established software application called Generic Importer and Exporter (GENIE) that has been used in NHS Scotland for the past two years in the context of SCI Diabetes, a national diabetes computing system (SCI-DC).

What is GENIE?

Automating the collection of information from different computer systems appears on the surface to be a simple task. Unfortunately this is not the case and the difficulty in achieving automation is illustrated by how uncommon this still is. The Scottish RCC of UK Biobank would address this problem through the use of a tool that greatly simplifies the task. Assuming a system can export its data to a text file (all the proposed feeder systems meet this requirement), GENIE can deal with the rest.

GENIE resides on a computer located on the site of the feeder computer system, that is, the general practice or SCI Store. At a time scheduled by GENIE, a snapshot of data is exported to a text file. GENIE compares this snapshot with the previous day's. Records that are new, modified or deleted (a small number each day) are then placed ready for transmission to UK Biobank. At a predefined time, these records are encrypted, compressed and transmitted to the site of the UK Biobank database server that is linked to the NHSnet. Should this transmission fail, records are queued for later transmission. On arrival at the central computer, the files are uncompressed and unencrypted. The data are then mapped, imported and merged with existing patient information. A schematic of the proposed infrastructure is shown in Figure 1. This is adapted from the system that is already in daily use for SCI-DC.

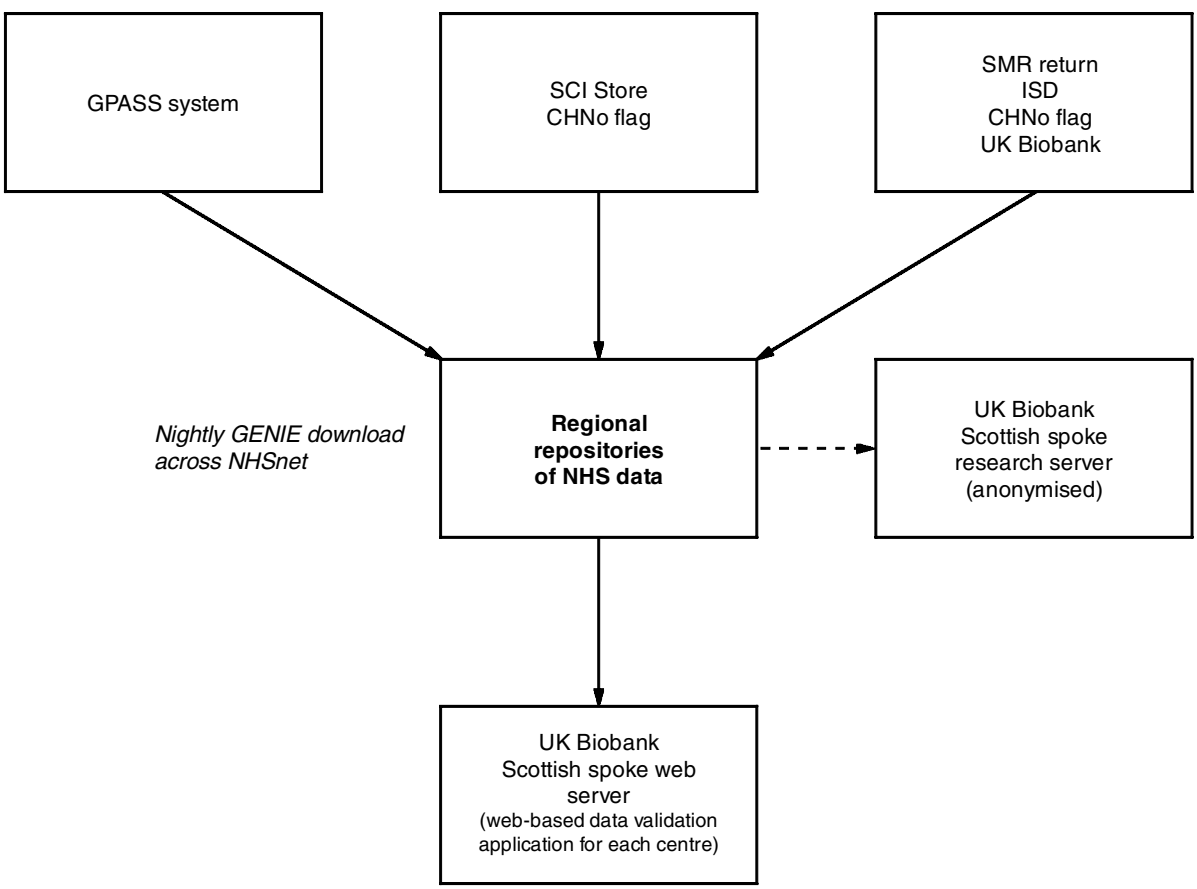

Figure 1 A schematic of the proposed automated structure

Quality assurance of data

Although there could be a single server for the Scottish RCC, the server could be partitioned so that only the research team that has recruited specific patients would be able to validate, supplement or quality-assure stored information via web-based technology. Users access the system via web browsers. A link to the national CHI database would ensure that the system knows who is alive or dead and which practice patients are registered with. The databases, web servers and back-up systems would be located in a secure machine room. Centrally, back-up resources, a GENIE server and a connection to NHSnet are required.

There are clear benefits of using this method to automate follow-up both during and after funding runs out for the main study. The benefits of a nightly download of prescribing history and events would give added value to the entire UK Biobank project. Thus we would gather information on event rates and prescribing rates that will be invaluable in powering genetic epidemiology and pharmacogenetic studies involving the whole 500 000 cohort. The datasets that we propose to use for automated record linkage are shown in Table 1.

We have proposed that the automated data management would be co-ordinated by the Health Informatics Centre at the University of Dundee and the Information and Statistics Division of the NHS in Scotland in a distributed model. Ultimately, a UK solution will be defined to which Scotland, like the other RCCs, will adhere. Some of these principles and mechanisms may be of interest to others in primary care who are using informatics tools and engaged in recruitment and follow-up of large numbers of patients.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the work of the senior programmers in developing these informatics tools, particularly Mr Douglas Boyle and Mr Neil McEwan. We would also like to thank Dr John Newton, the Chief Executive Officer of UK Biobank, and Professors Mike Pringle (University of Nottingham) and Richard Hobbs (University of Birmingham) for helpful comments on earlier drafts.

REFERENCES

1 Hapgood R, Shickle D and Kent A. *Consultation with Primary Care Health Professionals on the Proposed UK Population Biomedical Collection*. London: Medical Research Council and Wellcome Trust, 2001.
2 Bell J. The new genetics: the new genetics in clinical practice. *British Medical Journal* 1998;316:618–20.
3 Bumol TF and Watanabe AM. Genetic information, genomic technologies, and the future of drug discovery. *Journal of the American Medical Association* 2001;285:551–5.
4 Sullivan FM, Lewison G and Clarkson J. What Scottish primary care researchers are doing to recover their standing in the UK. *Health Bulletin* 2002;60:1–5.
5 The UK Biobank: www.ukbiobank.ac.uk/ (accessed 14/09/03).
6 Furness P. *A Briefing Document Regarding Sources, Acquisition, and Storage of Data for the Funders of the UK Biobank Study*. London: Department of Health/Medical Research Council/Wellcome Trust, 2003.
7 Horsfield P. Trends in data recording by general practice teams: an analysis of data extracted from clinical computer systems by the PRIMIS project. *Informatics in Primary Care* 2002;10:227–34.
8 Caldicott Committee. *Report on the Review of Patient-identifiable Information*. London: Department of Health,

Table 1 Datasets currently linked through the CHNo

1.1.1.1 Dataset	Description	Coverage	Dates
CHI dataset	NHS number (UPI) (includes date of birth/death)	Scotland wide, currently 5.6 million residents	1980–present
SCI Store	Regional repository of clinical data	In each health board region	2001–present
SMR1 record	Hospital admissions database ICD9/10 coded	Scotland wide	1980–present
GPASS	Primary care clinical and prescribing system, used by 85% of Scottish general practices	Scotland wide	1984–present
SMR 1, 2, 6, 10, 11, 12	Maternity, cancer, neonatal, child health databases	Scotland wide	1980–present

CHI: Community Health Index; UPI: unique patient identifier; SMR: Scottish Morbidity Record; ICD9: International Classification of Diseases version 9; ISD: Information Statistics Division, NHS Scotland

1997. www.doh.gov.uk/confiden/crep.htm (accessed 14/09/03).
9 Al-Shahi R and Warlow C. Using patient-identifiable data for observational research and audit. *British Medical Journal* 2000;321:1031–2.
10 Hewson B. Why the Human Rights Act matters to doctors. *British Medical Journal* 2000;321:780–1.
11 Strobl J, Cave E and Walley T. Data protection legislation: interpretation and barriers to research. *British Medical Journal* 2000;321:890–2.
12 Medical Research Council. *Personal Information in Medical Research*. London: Medical Research Council, 2000. www.mrc.ac.uk/pdf-pimr.pdf (accessed 14/09/03).
13 General Medical Council. *Confidentiality: protecting and providing information*. London: GMC Publications, 2000.
14 Warden J. Guardians to protect patient data. *British Medical Journal* 1999;318:284.
15 MacMahon S and Collins R. Reliable assessment of the effects of treatment on mortality and major morbidity. II: observational studies. *Lancet* 2001;357:455–62.
16 Thiru K, Hassey A and Sullivan F. Systematic review of scope and quality of electronic patient record data in primary care. *British Medical Journal* 2003;326:1070–4.
17 Marsden W, Duffy R and Sullivan F. *A Short Report on the Recruitment Potential of Two Scottish Primary Care Trust Areas to UK Biobank*. Dundee: TayRen/FresCo, 2002.
18 NHS Scotland IM&T Strategy: www.show.scot.nhs.uk/imt/ (accessed 14/09/03).

CONFLICTS OF INTEREST

None.

ADDRESS FOR CORRESPONDENCE

Professor Frank Sullivan
Tayside Centre for General Practice
Health Informatics Centre
University of Dundee
Kirsty Semple Way
Dundee DD2 4BF
UK
Tel: +44 (0)1382 632771
Fax: +44 (0)1382 633839
Email: f.m.sullivan@dundee.ac.uk

Accepted October 2003