ENSURING DATA PRIVACY IN BIOMEDICAL RESEARCH INVOLVING RECORD LINKAGES

Professor Chia Kee Seng
Department of Community, Occupational and Family Medicine
Yong Loo Lin School of Medicine
National University of Singapore

INTRODUCTION

Sources of personal information for biomedical research

1. There are many possible sources of personal information that can be used for biomedical research. These could range from information obtained through interviewing or testing a research subject or a patient, information submitted to a database or registry and information derived from tissues obtained from the research subject or patient. Broadly, we can consider all such information to be obtained in a research or non-research (usually clinical) context.
2. Within a research context, the collection of such data and tissues is subject to approval and review by ethical review committees and/or prevailing legislation/regulations. Informed consent is the accepted requirement for such collection.
3. The doctor-patient consultation is another major context where collated personal medical information could be used for biomedical research. Although obtaining informed consent is the preferred model, the research questions may not be apparent during the clinical consultation process. To return to the patient for informed consent may not be logistically practicable or be in the best interests of the patient.
4. Finally, data that is routinely collected or submitted to registries, public and private agencies may be of immense value for biomedical research. For example, data on death and emigration status is vital for follow-up studies. The disclosure

of such personal information is usually governed by existing national legislations or organizational regulations.

5. Such data is usually stored in the form of electronic records in databases managed by healthcare institutions, government and non-governmental registries as well as researchers. For the purpose of biomedical research, it may be necessary to link the records of individuals from multiple databases.

Value of record linkages

6. In 1994, Professor David Barker gave the Wellcome Foundation Lecture at the Royal Society of London titled ‘The fetal origins of adult disease’. His hypothesis was that the nutritional status of the embryo may program the developing fetus towards a higher risk of adult diseases like coronary heart disease, diabetes mellitus, stroke and hypertension. One of his earlier works was to trace the birth weights of 15,726 men and women born in Hertfordshire between 1911 and 1930 and their subsequent deaths from coronary heart disease till 1980s. Those with birth weights of less than 5.5 pounds and those who weighed less than 17 pounds at one year of age had the highest risk of coronary heart disease in adult life. Such findings were subsequently confirmed in several other countries.
7. The idea that adult health may reflect circumstances in childhood, or even earlier in life, is one that dates back many years. However, David Barker's group in the UK wondered if the effect of intrauterine programming extended to adult life. This simple but novel extrapolation of birth weight to subsequent coronary heart disease is commonly called “Barker’s hypothesis”. It generated much interest and controversy, with editorial comments ranging from the enthusiastic to the critical.
8. In much the same way, linking databases of patients suffering from a particular disease with the death registry helps doctors and medical researchers understand the natural history of diseases, identify prognostic factors as well as evaluate treatment strategies. At the national level, such linkages provide data for the evaluation of health care services and formulation of health care policies.

9. Although there is tremendous research and public health value in linking an individual's personal information, there is a need to respect and protect the privacy of the individual. In nearly all instances, researchers using the final dataset for analysis do not need to have the identity of specific individuals. The identity of the individual is only needed during record linkages and if the subjects need to be re-contacted. It is possible to develop systems that enable record linkages and re-contact, and at the same time protect the identity of individual subjects. Generally, such systems involve some degree of de-identification of the data collected from the subjects.

DE-IDENTIFICATION OF BIOMEDICAL RESEARCH DATA

Data that identifies an individual

10. Personal information about an individual with potential for research use can be divided into two groups:
    a. Personal identity data
    b. Research data
11. Personal identity data consists of data items that, singly or in combination, could potentially identify a specific individual. For example a person's name and unique personal identification number, which in Singapore is the National Registration Identity Card (NRIC) number, are considered personal identity data. Some would add ethnicity, date and place of birth, and gender. In rare situations, simple combinations like date of birth and diagnosis of an extremely rare condition may potentially reveal the identity of an individual. In other words, in rare situations, it may be possible to identify a specific individual from the research data. However, it is not necessary to invest vast resources to build a system that claims to be 'fail-safe' for such cases. Any system must balance the public interest against the protection of individuals, so that disproportionate costs are not involved in setting up and using a system that will not benefit or be relevant to the vast majority of cases. Systems must ultimately rest on positive

assumptions that the majority of users will want to honour the privacy and autonomy of their subjects, while putting in place a strong set of safeguards against potential misuse.

Terminologies

12. De-identification of biomedical research data can be defined as a process whereby personal identity data is separated from the research data. There have been many different terminologies for the concept of de-identification, such as anonymisation, pseudo-anonymisation, partial de-identification, etc.
13. Conceptually, it will be easier to have three levels of de-identification:
    a. Completely identifiable data: Personal identity data and research data are stored as single electronic or paper records. The data items may be coded or reversibly encrypted for confidentiality, but the personal identity and research data are physically linked within a single data table.
    b. Reversibly de-identified data: Personal identity data are separated from the research data. Each record in the research data is identified by a unique identifier such as a 'private unique identification number' (PUIN) which does not carry any personal identity information. The corresponding personal identity data is also identified by the PUIN which thus serves as a bridge between the two databases.
    c. Completely de-identified data: Personal identity and research data are de-linked in such a way that it is impossible to reconnect them and identify any individual from the research data.

De-identification in follow-up studies

14. The main characteristic of a follow-up or cohort study is the collection of data on predictive factors prior to the awaited outcome. In the famous Framingham study, subjects were recruited and information gathered on dietary and lifestyle factors as well as blood collected for measurement of cholesterol levels in individuals without heart disease. These subjects were followed-up for decades

during which, some of them developed coronary heart disease. Clinical trials follow the same design. A cohort of breast cancer patients is recruited and data on prognostic factors collected. These patients are randomly assigned to different treatment regimes and closely monitored for the outcomes of interest like recurrence, metastasis and death.

15. To maximize the value of data and tissues collected in such follow-up studies, the data should be managed as reversible de-identified data. Data managers should have in place a system for reversing the de-identification as new data on the same individual is obtained subsequent to the initial recruitment. However, the final datasets and samples sent to researchers should be completely de-identified.

16. A system for handling such reversible de-identified data should have the following characteristics:
    a. A trusted third party (TTP) with appropriate governance structure that holds the link between PUIN-personal identity data and
    b. A mechanism whereby the ground operations is partitioned such that no one is able to have all three sets of information: PUIN, personal identity data and research data.
    c. A mechanism of record linkage with external agencies such that they do not need to release completely identifiable data.

DE-IDENTIFICATION SYSTEM IN THE SINGAPORE CONSORTIUM OF COHORT STUDIES

The Singapore Consortium of Cohort Studies (SCCS)

17. The SCCS can serve as a specific illustration of the implementation of a system for maintaining privacy while allowing the collection and use of data from more than one source. The SCCS is an ambitious follow-up study by the National University of Singapore in collaboration with researchers from both healthcare clusters and A*STAR research institutes. The aim is to study how genetic and lifestyle factors influence each other in the risk of developing diseases of public health importance. It will establish two cohorts:

a. A multiethnic cohort of 250,000 normal healthy subjects for the study of their subsequent susceptibility to diseases like coronary heart disease, stroke and common cancers;

b. A multiethnic diabetic cohort of 25,000 type II diabetics for the study of diabetic complications.

18. The subjects and patients will be recruited with full informed consent and the entire project will be monitored by Institutional Review Boards. The Biomedical Research Council has provided initial funding for a 5-year pilot project.

19. In such cohorts, data on both genetic and lifestyle factors is needed. Genetic data of interest (germline mutations) will not change with the onset of disease. However, lifestyle factors change significantly and may affect recall of past lifestyle habits. Hence, data on lifestyle must be collected prior to the onset of disease. Furthermore, blood specimens will have biomarkers that could be used to estimate exposure factors.

20. Many countries around the world, are establishing such cohorts. The UK Biobank, for example, aims to recruit 500,000 subjects and ‘will be a unique resource for ethical research into genetic and environmental factors that impact on human health and disease, to improve the health of future generations.’ Similar efforts are seen in Sweden, US, China, Malaysia, South Korea and Japan. Unlike most of these countries, Singapore will provide a multi-ethnic cohort that has undergone rapid economic development resulting in dramatic changes in lifestyle factors. This combination of multi-ethnicity and rapid change is a powerful setting for discovering significant gene-environment interactions.

Maintaining data privacy in the SCCS

21. Recruitment of subjects for the cohort of normal healthy individuals will be done by field workers in the community setting. The field worker will therefore have the personal identity data and the questionnaire data. Each subject will also be identified at this stage using a unique study number (SN). The personal identity data and the questionnaire data will be coded, encrypted and kept in separate databases in the interviewer’s computer. At the end of each working day, the

databases are uploaded to the servers and the data in the interviewer's computer erased permanently. The questionnaire data will be sent to the research database while the personal identity data is sent to a separate database (Figure 1).

22. The research database (without the personal identity data) will be managed by SCCS staff from the NUS. The personal identity database will be managed by a Data Privacy Framework (DPF) Office under A*STAR. The DPF Office functions as the TTP and creates and maintains the unique PUIN. When the DPF Office receives the personal identity of a subject from the field workers, a PUIN is generated (if the subject has not been previously recruited) or the existing PUIN is retrieved.
23. When the SCCS Office receives the research data from the field worker, it will send the SN to the DPF Office, which will return the PUIN. The research data in the SCCS database are now tagged with this PUIN.
24. The subject is invited to a clinic for examination and donation of blood specimen. At the clinic, the NRIC is sent to the DPF Office which returns the original study number (SN). This study number will be used to track all the clinical data and specimens collected at the clinic. The clinical data is uploaded to the SCCS server directly. This clinical data is linked to the questionnaire data using the SN. The specimen is sent to the Singapore Tissue Network (STN) which is a nation-wide repository of biological specimens for research purposes (figure 2).
25. When the STN receives the samples, the SN will again be sent to the DPF Office which in turn will return the PUIN. Samples in STN will then subsequently be identified using the PUIN (figure 3).
26. In this system, the SCCS maintains an effective partition between different operations. No one will be in possession of all three sets of information: PUIN, personal identity data and research data.

Maintaining data privacy in electronic record linkages with external agencies

27. Over the years of follow-up, the SCCS subjects will be revisited for additional information. However, it may not be desirable to obtain information on the occurrence of certain outcomes (e.g. cancers and deaths) directly from the

subjects or their relatives. With electronic capture of such occurrences, it is possible to perform electronic record linkages with the respective registries.

28. For example, if a researcher needs information on episodes of coronary heart disease and death from coronary heart disease among the diabetics in the cohort, it is possible to obtain such information through electronic record linkages with both the Singapore Myocardial Infarction Registry (SMIR) and the Registry of Births and Death (RBD). However, a system must be in place to ensure that the privacy of the individual subjects is protected.

29. Following approval by the IRB for electronic record linkages with RBD and SMIR, the SCCS sends a listing of PUIN of all diabetics to the DPF Office. The DPF Office retrieves the NRIC of these subjects and creates a new number that is used only once (Nonce: number used only once). The DPF Office sends to SMIR and RDB the NRIC-Nonce listing. At the same time, DPF Office will also send to SCCS the PUIN-Nonce listing. The DPF Office will only hold the PUIN-NRIC-Nonce listing for a short duration. The RBD and SMIR will match the NRIC with their databases, identify the subjects with coronary heart disease, extract the necessary data items, and then remove the NRIC. Each individual with a coronary heart disease event will now be identified by the Nonce. The RBD and SMIR will send the listing of Nonce with the necessary additional data to SCCS. The SCCS can now link the new data on coronary heat disease to the diabetics using the Nonce. The necessary dataset can then be sent to the researcher without the PUIN.

30. This system will allow electronic record linkages with external agencies without revealing the identity of the subjects. Furthermore, the PUIN need not be sent to external agencies or researchers. The DPF Office will also not receive new data which has the potential for identifying individuals. New data will be sent to the SCCS without the personal identity data.

31. The entire operations of the DPF Office will be computerized and require minimal human intervention. The creation of Nonce, sending of listings to various bodies could be done automatically. The governance of the DPF Office

can be further strengthened by an independent Oversight Committee that will approve requests for electronic linkages as well as audit the operations.

CONCLUSION

32. There is tremendous research and public health value in linking an individual's personal information from various sources. In nearly all instances, researchers using the final dataset for analysis do not need to have the identity of specific individuals.
33. The model proposed here, though complex is a careful marriage of efficiency and privacy. It is easy to go overboard one way, at the expense of the other, so international best practices have been carefully studied, along with conditions and settings peculiar to the Singapore biomedical research scene. The model provides an example of what can be done to enable important research that will provide great benefit in terms of helping prevent diseases and their complications by identifying risk factors.

Figure 1: SCCS operations – fieldwork

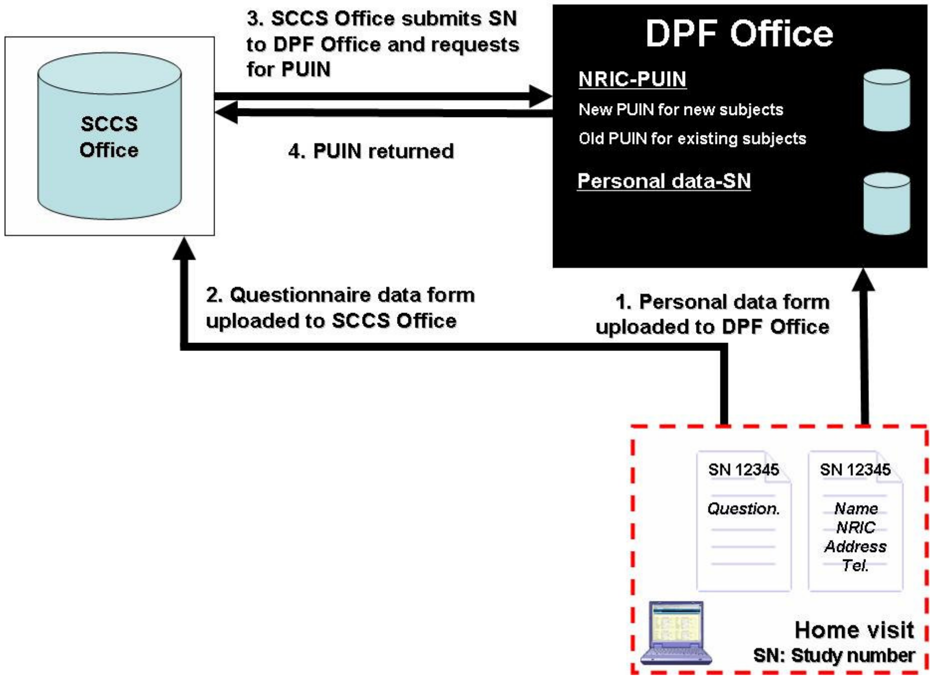

3. SCCS Office submits SN to DPF Office and requests for PUIN
4. PUIN returned
SCCS Office
DPF Office
NRIC-PUIN
New PUIN for new subjects
Old PUIN for existing subjects
Personal data-SN
2. Questionnaire data form uploaded to SCCS Office
1. Personal data form uploaded to DPF Office
SN 12345
Question.
SN 12345
Name
NRIC
Address
Tel.
Home visit
SN: Study number

Figure 2: SCCS operations – clinic visit

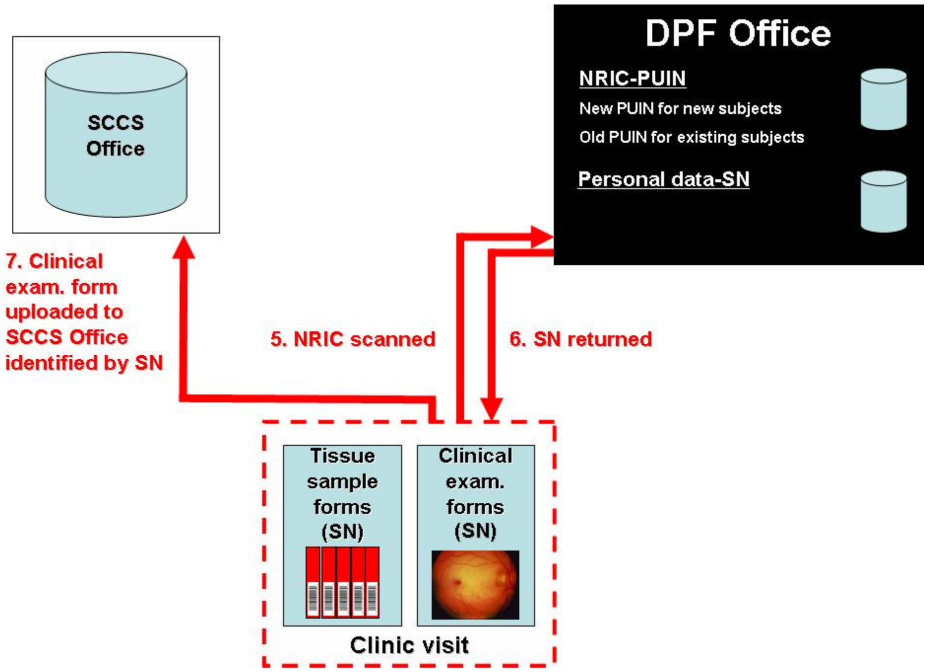

SCCS Office
DPF Office
NRIC-PUIN
New PUIN for new subjects
Old PUIN for existing subjects
Personal data-SN
7. Clinical exam. form uploaded to SCCS Office identified by SN
5. NRIC scanned
6. SN returned
Tissue sample forms (SN)
Clinical exam. forms (SN)
Clinic visit

Figure 3: SCCS operations – STN

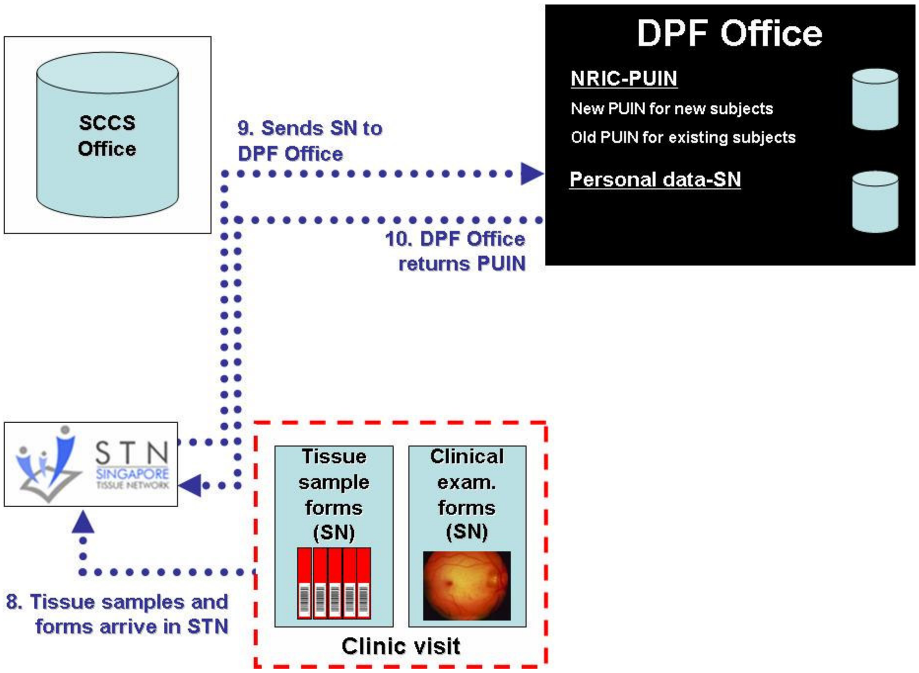
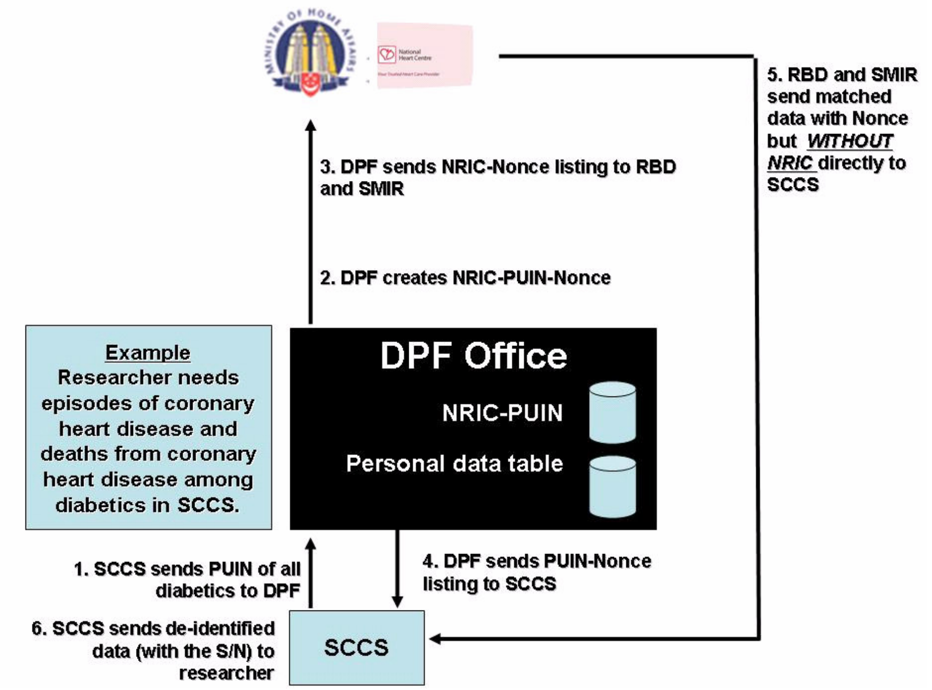

DPF Office
NRIC-PUIN
New PUIN for new subjects
Old PUIN for existing subjects
Personal data-SN
SCCS Office
9. Sends SN to DPF Office
10. DPF Office returns PUIN
STN SINGAPORE TISSUE NETWORK
Tissue sample forms (SN)
Clinical exam. forms (SN)
Clinic visit
8. Tissue samples and forms arrive in STN

Figure 4: SCCS operations – electronic record linkages


5. RBD and SMIR send matched data with Nonce but ***WITHOUT*** ***NRIC*** directly to SCCS
3. DPF sends NRIC-Nonce listing to RBD and SMIR
2. DPF creates NRIC-PUIN-Nonce
Example
Researcher needs episodes of coronary heart disease and deaths from coronary heart disease among diabetics in SCCS.
DPF Office
NRIC-PUIN
Personal data table
1. SCCS sends PUIN of all diabetics to DPF
4. DPF sends PUIN-Nonce listing to SCCS
6. SCCS sends de-identified data (with the S/N) to researcher
SCCS