NCIN
national cancer intelligence network
Using information to improve quality & choice

Building an e-health research infrastructure for cancer

A consultation

NCRI
National Cancer Research Institute

Consultation on building an e-health research infrastructure for cancer

Executive summary

Over the past three years the National Cancer Intelligence Network (NCIN) has built on the long history of cancer registries supporting research. The National Cancer Data Repository has allowed easier access to England-wide data, while linkages to Hospital Episode Statistics and other sources have enabled novel work. An ongoing modernisation of the cancer registration service in England, combined with new datasets now either mandated for collection or in the process of mandation, is making a step change in the data available. Together, these developments will provide a timely, quality assured, patient-level longitudinal record of cancer care. This will be a unique resource for research.

It is therefore timely to give renewed consideration to the ways in which the cancer registration system can enable research. This document describes a number of ways in which NCIN could potentially facilitate or provide infrastructure to enable cancer research using linked health and health-related datasets. They cannot all be taken forward at once, and any of these ideas would need more detailed scoping and assessment of resource need as a next step. Other initiatives, in particular the recently announced Clinical Practice Research Datalink (CPRD), are also developing plans for research infrastructure and it will be important to co-ordinate activity. The purpose of this consultation is to assess demand for the services and guidance indicated and to assist in determining priorities. NCIN could offer:

Proposal 1: A safe haven facility to enable research access to potentially identifiable data from the English cancer registration service in a secure, controlled environment. All outputs removed from the safe haven would be checked against anonymity criteria before release.

Proposal 2: A linkage service to allow combination of different datasets for use in research. Researchers with the appropriate consent could receive data from the cancer registration service to supplement a study dataset. Other researchers could receive access to the de-identified product of linking one or more existing datasets to cancer registry data.

Proposal 3: A notification service for new cancer diagnoses and other events. Investigators could register study participants (with their consent) with the cancer registration service and allow them to be regularly informed of new diagnoses and other events.

Proposal 4: A hosting service for study data. This would provide investigators with a secure environment in which to store and analyse their study data together with routinely collected information from the cancer registration service.

Proposal 5: Support for the planning and recruitment of studies. This would involve mechanisms for estimating trial feasibility and the likely recruitment rates of different centres, and for notifying clinical teams when patients are eligible for a particular study or clinical trial. (Potentially a joint project with the NIHR Cancer Research Network (NCRN) and others).

Proposal 6: A research support service. This would advise users on the availability of data, the requirements for access, and would manage requests and provide access.

Proposal 7: Guidance on guidance on permissions for data access and linkage. This would involve providing investigators with guidance on consent wording for using record linkage and notification services. (Potentially joint projects with NRES and NIGB and others).

If there is interest in developing any of these proposals NCIN will discuss with National Cancer Research Institute partners, NCRN and CPRD how best to develop such infrastructure. To inform these discussions, which will inevitably include consideration of cost, it would be especially helpful to have information on:

- Who would use these resources and the likely level of demand;
- What types of research or analysis would become possible for the first time;
- Other benefits such as enabling more timely access to data;
- Priorities for the various potential resources.

Comments on both these issues and more generally on the proposed infrastructure would be welcome. Please send these to research@ncin.org.uk by 9 Jan 2012.

Contents

Introduction

It is recognised in both the public[1,2] and commercial sectors[3] that the NHS provides the UK with a unique opportunity for health research. The existence of a universal healthcare system, providing services to around 51m people in England[4] and maintaining comprehensive records, is a hugely valuable opportunity for research, including research which makes use of electronic health records (e-health research). However, it is also recognised that there are weaknesses in NHS data as a research resource. In secondary care the readily available datasets are collected for administrative purposes, with concerns around the quality of coding. Individual datasets cover only a specific part of the care pathway and therefore require linkage to investigate many research questions, while control of access to datasets is fragmented across different organisations. In primary care the availability of data for research is patchy, with datasets being largely anonymous, hindering linkage (although notable progress has been made by the General Practice Research Database[5], GPRD), and only a small proportion of the population is covered by any given primary care dataset.

Meanwhile, the importance of access to large-scale health datasets is growing. A particular challenge for the coming decade will be the increasing stratification of treatments and their tailoring to much smaller subsets of patients[6]. The development of these stratified treatments will depend in part on the availability of large, population-based and high resolution datasets of clinical information on individual patients[7]. To begin to address these challenges, the Government have announced the creation of a new service through combining the GPRD with the Health Research Support Service established by the Research Capability Programme[8]. The new Clinical Practice Research Datalink (CPRD) will develop a range of capabilities for both observational and interventional research over the next four years, with the overall aim of strengthening the international competitiveness of UK life sciences research.

In cancer, it is already clear that treatment needs stratification if it is to deliver the maximum benefit to each patient[9], and to achieve this requires data linkage across the complex, multi-modality and multi-organisational treatment pathway. Allocating patients to the correct treatment, monitoring cancer care and assessing outcomes can only be achieved by linking data from disparate systems.

Cancer registries have existed for many years and ensure robust ascertainment of cancer patients across the care pathway by accessing data feeds from many different local and national sources. The UK benefits from population based cancer registration through a system of eleven cancer registries (eight in England and one each in Scotland, Wales and N. Ireland)[10]. Cancer registries are authorised by Parliament to process confidential information about cancer patients for purposes which include medical research[11]. All of this means that cancer is well positioned to act as an exemplar for other disease areas.

The National Cancer Intelligence Network (NCIN) was created in 2008 to improve standards of cancer care and clinical outcomes by improving and using information about cancer patients[12]. Established under the umbrella of the National Cancer Research Institute, one of its core objectives is to enable new research opportunities through the provision of a world leading e-health research infrastructure in cancer. The NCIN has the national system of cancer registries as its core and has championed a national cancer information strategy which draws on the skills, information flows, quality assurance processes and information governance framework of the English cancer registries. Cancer registration has traditionally been seen by researchers as a source of epidemiological data on

incidence, mortality and survival. The strength of the UK's cancer registration system has supported the development of world-leading research groups. The much richer dataset now being put together will substantially enhance the research potential.

Over the last five years the remit of the regional English cancer registries has been expanded to include collection of a much more extensive range of cancer-related information on individual cases and the development of specialist expertise in particular disease types. The NCIN is now managing a major programme of modernisation across all eight English cancer registries, which will see the creation of a unified national cancer registration service by the end of 2012. This service will ensure that common standards and working practices are applied to data extraction, linkage and quality assurance to both national feeds and a range of local sources. These sources will be combined in a patient-level, structured and quality assured cancer dataset that holds at a minimum all 760 data items of the new Clinical and Services Outcome Dataset[13]. Work is also underway to establish a patient-level linkage between the national cancer screening programmes and the national cancer registration system. The unified service will provide a dynamic view of patient care and cancer diagnosis. Automation and the use of distributed expertise across the regional registry offices allows data collection and reporting to be in near-real-time – data collected one month are available the next. The historical data from the existing regional cancer registries is also fully integrated to provide a longitudinal linked dataset stretching back for almost 50 years in some geographical areas.

This modernisation programme will deliver one of the most sophisticated health datasets and clinical processing environments in the world by early 2013. Epidemiological and health services research may be performed solely on the dataset[14] or the data may be used to complement and supplement other research datasets (such as clinical trials, cohort studies or tissue resources).

Data are currently available to researchers who must apply to one of the regional cancer registries under rules set by the UK Association of Cancer Registries (UKACR) [15] and agreed with Ethics and Confidentiality Committee of the National Information Governance Board for Health and Social Care (NIGB ECC). While cancer registries have a long history of supporting and undertaking research, this remains a relatively *ad hoc* process managed by the staff alongside their core NHS work, which they must prioritise. Researchers may also apply for access via the GPRD. This route provides access to anonymised data relating to patients in the GPRD (approximately half of GPRD practices take part in the linkage scheme, covering roughly 4% of the population).

There is the potential to build on this unique dataset to provide the UK with a world-leading capability for health research in cancer. The combination of a population-based, high resolution, clinical dataset together with an infrastructure designed to facilitate research on these data would be of huge value to cancer research internationally. The remainder of this paper sets out a vision of what this infrastructure might look like, and actions that the NCIN could take to achieve it. We anticipate that the capabilities described could be built in a stepwise fashion, with demand, feasibility and benefit proved at each stage. Given existing links with the GPRD and the Research Capability Programme, we expect that data from the cancer registries will, in due course, form an integral part of the new CPRD service. Any infrastructure developed in cancer would need to complement rather than duplicate the existing and developing national research infrastructure, including the CPRD. However, such infrastructure could act as a test case and exemplar for services in other disease areas and contribute to the research capability of Public Health England as this emerges.

A vision for e-health research in cancer[16]

An outline set of requirements for an e-health research infrastructure are proposed in table 1. These are not cancer specific, but the unique legal support afforded to the cancer registries means that the underlying data are more readily available for cancer patients. Building on existing strengths in cancer would both enable research on a major disease and develop techniques and capacity which could in future be applied in other disease areas through links with the CPRD and other initiatives.

Table 1. Requirements for an e-health research infrastructure.

Requirement	Capability
Longitudinal patient records are centrally accessible	Obtain and link core datasets for patients directly from hospital trusts and from NHS administrative data sources.
Access to aggregated datasets for use in research	Timely extraction of anonymised datasets in response to appropriate requests. Strict controls to ensure that individual requests or combinations of requests are not disclosive.
Authorised users can access potentially identifiable datasets in a controlled environment.	***Safe haven*** where authorised researchers can analyse potentially identifiable data, with all outputs checked against anonymity criteria before release.
Studies (trials, cohorts, tissue resources, etc.) can be supplemented with additional patient data. **Health datasets can be linked to identifiable health, social care or other databases, with an anonymised result provided to an authorised investigator.**	***Linkage service*** to match records based on identifiers. Able to return either pseudonymised data associated with a study ID or an anonymised dataset resulting from the linkage of two identifiable datasets.
Investigators running clinical trials or other studies can receive notification of new diagnoses and details of treatment and outcomes.	***Notification service***, building on the flagging service currently provided by the NHS Information Centre and able to provide data on a wider range of events.
Investigators can store their study data securely within a national infrastructure and receive access to a linked dataset containing both research and routinely collected items.	***Data hosting service*** to hold research datasets alongside routine data in a secure environment. Able to provide study owners and, subject to permissions, other investigators with access to anonymised, linked datasets with combining routine data with that collected in the study.
Investigators planning studies can discover how many patients meet the eligibility criteria and, for trial recruitment, where they present. **Where appropriate, clinical teams can be notified when a patient is eligible to take part in a study.**	***Research planning service*** to allow feasibility counts, estimation of recruitment rates based on eligibility criteria and identification of centres treating large numbers of eligible patients. ***Participant recruitment service*** to notify clinical teams when one of their patients is eligible for a particular study based on its eligibility critieria.
Potential users are aware of available data / services. **Ethical, legal and regulatory requirements for use are clear.**	***Research support service*** to advise users on the availability of data, the requirements for access and to manage requests and provide access. Would work with regulators to establish clear controls for the use of the various services.

Proposal 1: A safe haven to enable access and protect confidentiality

Extracts of tabular data from the cancer registries are relatively freely available for research and other purposes, providing the numbers of cases or underlying populations are sufficiently large. Although effectively anonymised in isolation, there is a risk that such datasets become potentially identifiable in combination (for example, the difference between two datasets with overlapping geographical or temporal coverage may leave a 'sliver' covering a small population). There is considerable scope for informatics research in disclosure control techniques to mitigate this and other risks but in general research users are able to access the datasets that they need.

A more significant barrier for research is the need to access identifiable or potentially identifiable datasets. Potentially identifiable data refer to patient level data or datasets relating to small areas or to low numbers of patients in small populations[15]). Access to identifiable datasets is often time consuming for the researcher, who must obtain the appropriate consent or special permission. Few investigators need identifiable data for purposes other than linkage; some infrastructural possibilities to further reduce the need to access such data are outlined below.

A much larger number of research projects require access to record level data or data relating to small populations. Access to these potentially identifiable data requires the director of the relevant cancer registry to assess the risk of disclosure and the purposes for which the data are being released. In some cases data may need to be pre-processed in the registry to remove the risk of disclosure. All of this takes time and delays research, particularly since the process of agreeing a dataset for release is frequently iterative.

Another model which has been applied is for researchers to work within the registry environment under an NHS contract or strict governance controls. This allows them access to potentially identifiable data in a secure environment (from which the data may not be removed). Setting up such an arrangement is inherently time consuming and organisational policies on the granting of access or contracts vary. This arrangement does, however, introduce the idea of a 'safe haven' for access to data that are not directly disclosive but which may become so when combined with other information. By ensuring that data are only processed within the safe haven and applying disclosure control to all outputs which leave the safe haven, the risk of disclosure can be reduced significantly.

The concept of safe havens was supported by the Data Sharing Review undertaken by Sir Richard Thomas and Sir Mark Walport in 2008[17]. Examples of existing safe havens include the ONS Virtual Microdata Laboratory[18], the UK Data Archive's Secure Data Service (SDS)[19] and the HRMC datalab[20]. Such a safe haven could be physical, requiring researchers to come to a particular location to access data, or virtual, with technological controls applied to achieve the same level of security. The SDS provides an existing example of a virtual safe haven. Approved and trained researchers are able to access and analyse data either on their own computer or in an approved secure room within their institution. No data can be downloaded and only non-disclosive outputs (papers and presentations) may be removed from the environment after review by SDS staff. A virtual model poses technical challenges and may be expensive but is appealing due to its scalability, running costs, and the low barriers to research use. On the other hand the cancer registration system offers an existing infrastructure of regional offices which could provide physical access for a relatively lower cost.

Potential action: A safe haven facility to enable research access to potentially identifiable data from the English cancer registration service in a secure, controlled environment.

Proposal 2: Enhancing studies through linkage of datasets

Using routine healthcare data to supplement research datasets can help improve completeness by providing missing data items, reduce the cost of follow up, and allow novel studies not possible with the original dataset. For example, collections of tissue samples need high quality data about the donor, their medical history and their treatment. These data could be collected by research nurses but, where they already exist in the cancer registry, it is more effective to collect these once and make them available to a variety of authorised users. Likewise, the equivalent of almost 20% of new incident cancer cases are now enrolled in a clinical trial[21], a success which requires increasing resources to be devoted to follow up of trial participants. Much of the information collected during follow-up will also be held by the cancer registry. A link to the cancer registration dataset could reduce the cost of follow-up of these patients as well as provide information on patients who would otherwise be lost to follow up. Such a link can also allow a comparison of the trial population with the general population of cancer patients[22].

Where studies involving cancer patients hold consent to access health records such information can be supplied if capacity is available. Once the researcher has shown that appropriate consent is held to access the information and identifiers have been provided for the study participants, the relevant individuals can be identified in the cancer registry database and the requested data items can be provided accompanied by study identifiers. Conducting the linkage between trial and cancer registry datasets does require capacity and expertise within the cancer registries. To date this has been conducted as a research activity rather than a service, but there is little reason not to develop a capability to offer this routinely to researchers.

Potential action: A linkage service, which, subject to appropriate consent, would allow researchers to submit identifiers for participants in a study and receive data from the cancer registration service to supplement the existing study dataset.

The capacity to link data sources would provide further opportunities to support research on linked datasets. One case is where an investigator holds an environmental or organisational dataset that could be linked to the cancer register (for example pollution data by area or information about NHS organisations commissioning or providing care). Alternatively, the investigator may wish to link an existing dataset about cancer patients held by a third party to information from the cancer register. While no identifiers are required for the research, identifiable or potentiality identifiable data are required for the linkage. Here the cancer registry would receive the details from the data owner (under the permissions which allow processing of data about the diagnosis and treatment of cancer in support of medical research), conduct the linkage, and make a de-identified dataset available to the research in a safe haven environment.

Potential action: A service which would link datasets relating to cancer patients or their environment to the cancer register and provide researchers access to the linked, de-identified data in a controlled environment.

Proposal 3: Notifying investigators of new cancer diagnoses and other events

Once in place, a natural extension of a linkage service is notification of new cancer cases among participants in a research study. There is already a well-established flagging mechanism for cancer cases, provided by the Medical Research Information Service at the NHS Information Centre[23]. This provides basic details of the diagnosis and where it was registered, allowing additional information

to be sought if required. At present not all studies make use of this service and funders might give consideration to supporting greater use of this service.

A notification system linked directly to the cancer registry system could supplement this by not only providing greater detail about the diagnosis but also allowing regular updates of any events relevant to the study. For example, although limited at present, information on progression, treatment and outcomes will be routinely available in future.

Potential action: A notification service, which would allow investigators to register study participants (with their consent) with the cancer registration service and allow them to be regularly informed of new diagnoses and other events.

Proposal 4: Providing a hosting service for research data

The discussion above supposes that investigators will hold data and the identifiers required to link participants to routinely collected data in their own systems. It is also possible given the flexibility and relative ease of extending the dataset collected by the cancer registration service that additional data items for research studies could be held within the registry. This is the model that has been adopted by the Cancer Research UK Stratified Medicine Initiative, which is using cancer registry systems to centrally host data collected in participating centres.

Such a model would provide a secure environment within which study data could be hosted. Authorised investigators would be able to access study data, together with routinely collected data as required, perhaps within a safe haven environment. This would reduce duplication in the collection of data items (as items routinely collected would already be available), allow for study data to be retained in the long term, and could facilitate data sharing if study data were available for release together with the other data held within the cancer registry after an agreed period.

Potential action: A hosting service for study data, which would provide investigators with a secure environment in which to store and analyse their study data together with routinely collected information from the cancer registration service.

Proposal 5: Supporting the planning and recruitment of studies

Although there has been great success in recruiting cancer patients into clinical trials, meeting the targets set for both the rate of enrolment and overall number of patients enrolled remains a challenge[1,21]. Cancer registries regularly provide feasibility counts to assist with the planning of observational studies. There is scope for cancer registry data to help plan trials, both in terms of identifying the centres where eligible patients present and estimating more accurately how long trial recruitment will take. Provided the eligibility trial criteria are available within the data collected by the cancer registries, then this dataset can be queried to identify how many patients are eligible and where they present.

Potential action: A mechanism for estimating trial feasibility and the likely recruitment rates of different centres (potentially a joint project with the NIHR Cancer Research Network (NCRN)).

In some cases (for example research on patient experience or involving cancer survivors) cancer registries may also be able to assist with identifying patients who are eligible to take part in research. Data will not be timely enough to support most interventional research but where a delay between diagnosis and recruitment is acceptable and eligibility criteria that can be expressed in terms of the information collected by the cancer registries, it would be possible to identify eligible

patients for a particular trial. Registries could then notify the relevant clinical team (thus avoiding breaching confidentiality) so that patients can be offered the opportunity to participate in the research.

Potential action: A mechanism for notifying clinical teams when patients are eligible for a particular study or clinical trial (potentially jointly with NCRN and others).

Proposal 6: Providing additional support to research users of data

Potential users of the types of infrastructure described above would need support to understand the data available, the potential routes and models of access and the permissions required. A more formal research support service could be established to manage the infrastructure and support users in gaining access to it. The individuals providing such support would be integrated with the cancer registration service to ensure an understanding of the data available and its potential uses. A central tracking system for applications to use data could be established, with details of approved applications made publically available to both ensure transparency in the uses of data collected by cancer registries and provide a measure of the usefulness of the resource.

Potential action: A research support service that would advise users on the availability of data, the requirements for access, and would manage requests and provide access.

Proposal 7: Providing guidance on permissions for data access and linkage

A major consideration for such a research support service would be ensuring that the correct permissions are in place for all of the research activity taking place. Difficulties do occur when the consent given for a study does not clearly specify ongoing access to a patient's medical records for follow up or if the study does not hold the necessary identifiers. Whilst NCIN cannot directly solve either of these problems, there is scope to work with the relevant organisations to provide information and guidance to improve this in future studies.

Potential action: Clear guidance on forms of consent wording that are acceptable for linkage to records of this type (potentially a joint project with the National Research Ethics Service, the National Information Governance Board and others).

Potential action: Guidance to principal investigators and trials units that identifiers may be retained if participants give consent and these are required for purposes of linkage (potentially a joint project with trial sponsors, the Medicines and Healthcare products Regulatory Agency, the NCRN and others).

For trials or other studies involving only cancer patients cancer registries have permission to receive details about these individuals. However, where studies involve individuals who have not been diagnosed with a cancer, then a linkage or notification service would need to operate under clear consent from the patient that their data could be provided to the cancer registry for the purpose of linkage or notification in the event of a cancer diagnosis.

Potential action: Example forms of words for participants in research to allow their identifying information to be stored and processed for purposes of linkage or notification of new cancer cases.

Finally, there will also be occasions when investigators wish to link datasets for which the appropriate consent is not held and which relate to patients without a diagnosis of cancer. Here special permission would still need to be sought for registries to link these data. Where such permission is obtained, it should be possible to conduct the linkage within the cancer registry and make the de-identified data available within a safe haven very much as described above.

References

[1] HM Treasury and Department for Business Innovation and Skills. (2011) The Plan for Growth. Available from http://www.bis.gov.uk/policies/growth/the-plan-for-growth

[2] UKCRC R&D Advisory Group to Connecting for Health. (2007) Report of Research Simulations. Available from http://www.ukcrc.org/infrastructure/e-healthresearch/advisorygroupcfh/

[3] Association of the British Pharmaceutical Industry. (2011) The Vision for Real World Data – Harnessing the Opportunities in the UK. Available from http://www.abpi.org.uk/our-work/library/industry/Pages/Vision-for-Real-World-Data.aspx

[4] NHS Choices: About the NHS (http://www.nhs.uk/NHSEngland/thenhs/about/Pages/overview.aspx). Last updated 04/12/2009.

[5] See http://www.gprd.com for details of the General Practice Research Database.

[6] Trusheim MR et al. (2007) Stratified medicine: strategic and economic implications of combining drugs and clinical biomarkers. Nature Reviews Drug Discovery. 6:287-293. doi:10.1038/nrd2251.

[7] European Commission, DG Research - Brussels. (2010) Workshop report: "Stratification biomarkers in personalised medicine". Available from http://ec.europa.eu/research/health/pdf/biomarkers-for-patient-stratification_en.pdf

[8] See http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_130835 for the announcement of the Clinical Practice Research Datalink.

[9] See http://science.cancerresearchuk.org/research/research-strategy/our-progress/stratified-medicine-programme/ for information on Cancer Research UK's Stratified Medicine Programme.

[10] See http://www.ukacr.org/registration-organisation for details of the UK's cancer registration system.

[11] Health Service (Control of Patient Information) Regulations 2002

[12] See http://www.ncin.org.uk/ for details of the National Cancer Intelligence Network.

[13] See http://www.ncin.org.uk/collecting_and_using_data/data_collection/ncds.aspx for full details of the Clinical and Services Outcome Dataset.

[14] See http://www.ncin.org.uk/publications/peerreviewed.aspx for publications using data from NCIN.

[15] See http://www.ukacr.org/content/approved-ukacr-policies for UK Association of Cancer Registries policies.

[16] This paper focuses on England since health services are devolved and vary between the UK countries, making it natural to consider each individually, and England provides the largest population yet has made least progress in record linkage. (This is likely to change with the announcement of CPRD). Major studies or trials will frequently be UK wide; the potential for harmonisation of approaches between the UK nations should be explored to facilitate this.

[17] See http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Personal-information/Data-Sharing-Review/index.htm for key recommendations of the Data Sharing Review.

[18] See http://www.ons.gov.uk/ons/about-ons/who-we-are/services/vml/index.html for details of the Virtual Microdata Laboratory.

[19] See http://securedata.data-archive.ac.uk/ for details of the Secure Data Service.

[20] See http://www.hmrc.gov.uk/datalab/ for details of the HMRC Datalab.

[21] NIHR Cancer Research Network. (2011) Annual Report 2010-11. Available from http://ncrndev.org.uk/downloads/MiscDocs/NCRN%20Annual%20Report%202010-11.pdf

[22] Morris E.J. et al. (2010), Comparison of treatment and outcome information between a clinical trial and the National Cancer Data Repository. British Journal of Surgery, doi: 10.1002/bjs.7295.

[23] See http://www.ic.nhs.uk/mris for details of the Medical Research Information Service.

November 2011

National Cancer Intelligence Network Coordinating Centre
18th Floor, Portland House
Bressenden Place
London, SW1E 5RS

enquiries@ncin.org.uk

www.ncin.org.uk

NCIN
national cancer intelligence network
Using information to improve quality & choice