Linking investigators

A centralized linking-facility for data sharing and coordination of samples in tissue banks

Neil R. Smalheiser

Many multifactorial diseases, such as various forms of cancer, diabetes, coronary heart disease and schizophrenia, still pose a considerable challenge for those researchers intent on understanding their causes. Some progress has been made in identifying the genes that contribute to such diseases and the availability of the sequence of the human genome holds further potential. However, scientists are still far from pinpointing the exact genetic causes of most major diseases that now plague humans in the developed world. Clinicians and epidemiologists have amassed a wealth of non-genetic information, notably lifestyle and environmental factors, and with the rise of genomics, proteomics and imaging methods, banks holding tissue samples from patients have become an increasingly important and precious resource in medical research. However, the full value of all these efforts can only be realized by data sharing.

A major problem is the fact that the data generated by researchers around the world are not easily available to those working in a related area. Although it has become considerably easier to identify relevant scientific publications, and to access gene or protein sequence data and analytical software, this unfortunately does not extend to primary data from studies and tissue banks. Indeed, scientists could gain valuable information if they had access to these primary data or to samples from their colleagues' studies. Current approaches, namely the pooling of diverse data in central databases and the formation of consortia between investigators, are fraught with the practical problems of data management, ownership of data and confidentiality. An alternative, as proposed in this Viewpoint, would be a centralized linking facility to issue anonymous 'linking numbers'. These would allow independent and autonomous research groups studying samples from the same subjects to be aware of each other, and to form voluntary and temporary collaborations for research on a specific problem. Without incurring any confidentiality problems, such a facility could enhance progress, as it would allow scientists to identify related studies and to pool and analyse their data.

Scientists could gain valuable information if they had access to these primary data or to samples from their colleagues' studies

Clearly, this need to pool data from genetic and epidemiological studies has become increasingly important for research into human diseases. Moreover, the relevant variables for a multifactorial disease include demographic, clinical, developmental, psychosocial, environmental, genetic, epigenetic, anatomical, physiological and molecular factors, and are too numerous to be considered by any one investigator or to be controlled within any one study. To overcome these limitations, investigators have already established consortia to achieve adequate statistical power, as well as economy of scale (Torrey *et al*., 2000; Geschwind *et al*., 2001). However, there are practical problems with this approach. First, investigators are often reluctant to deposit data that they have not fully analysed into databanks shared with other scientists. Second, a central database must have access to patient identifiers, which raises problems of confidentiality. Third, a central database cannot be easily scaled-up to large numbers of investigators, and becomes increasingly difficult to maintain as more scientists submit their data. Furthermore, there are also strong arguments for investigators to remain autonomous and independent.

A major challenge for informatics is thus to link independent research groups in a manner that facilitates data sharing, data coordination and mining of pooled primary data. Here, I consider one particular aspect of this problem—the situation that occurs when different research groups obtain samples from tissue banks and perform studies on samples taken from the same individuals. Examples of such tissue banks include repositories of DNA and tumour specimens and post-mortem brain banks, as well as banks of plasma, serum and other blood components. In the past, such tissue banks have been primarily used for biochemical and histological studies, but have now attained even more importance with the development of genomic and proteomic methods. Tissue banks permit investigators to test hypotheses on a set of well-characterized subjects, and have the potential to accelerate research by allowing researchers to combine primary data from the same individuals, but from different studies. In addition, imaging archives (Van Horn *et al*., 2001) can also be regarded as equivalent to tissue banks, as they provide investigators with raw data for further study.

As outlined above, a centralized database where investigators could deposit their primary data would have several practical limitations. An alternative would be to link the data across numerous databases belonging to individual investigators. To facilitate this, I propose establishing a centralized linking facility that assigns a unique 'linking number' to each subject at the time of sample deposition. Let us suppose that such a facility is established for the benefit of any researcher who would like to donate samples to tissue banks or to carry out studies on such samples. At the time that an investigator donates a sample from a research subject to one or more cooperating tissue banks, the linking facility would issue a linking number both to the tissue banks and to the donating investigator. This number would be attached to all data relating to the individual, in both the tissue banks and the donating investigator's own database, including any new data that may be generated from that individual. In short, the linking number becomes an arbitrary and limited ID number for the subject, and is different from the notion of a universal identifier, which has received much attention in recent years (White & Messler, 1997). Although a tissue bank could use its own

Tissue banks have the promise of accelerating research by allowing researchers to combine primary data from the same individuals but from different studies

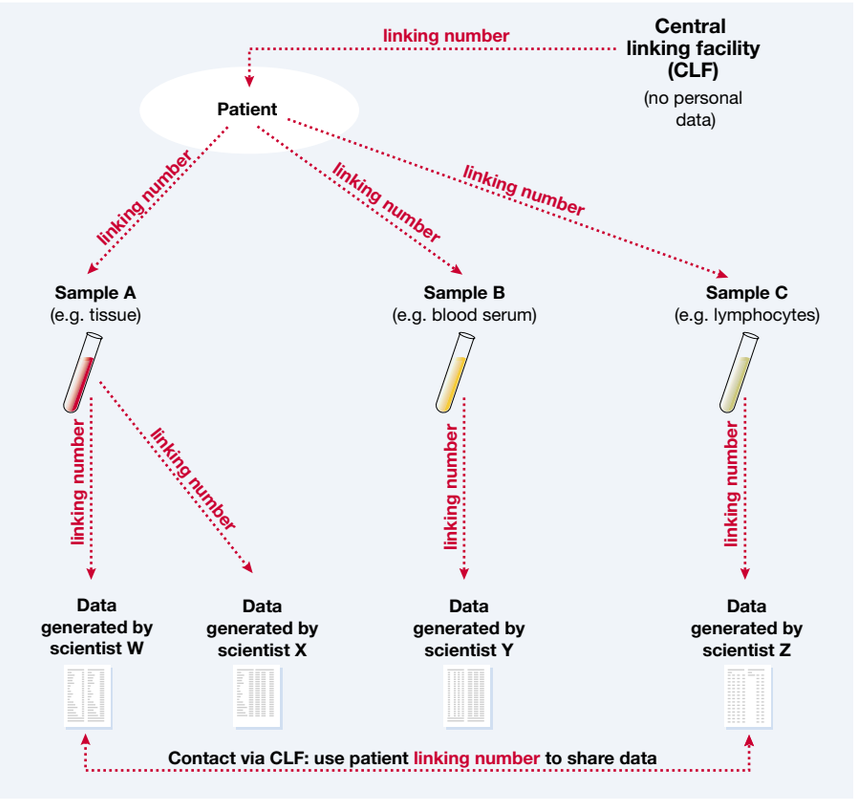
sample-coding numbers for the purpose of linking investigators, a separate linking facility is preferred if investigators donate material from the same subject to multiple tissue banks and/or data repositories; for instance, if lymphocytes are donated to a DNA bank and functional magnetic resonance imaging data from the same individual are donated to an image bank.

When other investigators receive samples from these linked tissue banks, they will also get the linking number that they are obliged to attach to all data they generate from the sample. The centralized linking facility will not have any identifying information, so there is no loss of confidentiality (Lavori *et al*., 2002). Moreover, there is no central database, so there is no need for investigators to deposit their research data. The linking facility will just keep track of which investigators have donated or received which samples, regardless of the tissue bank from which they were obtained, together with a brief description of their planned studies. This will allow different groups that have studied samples from the same individuals to locate one another, pool their results, and analyse them collaboratively, without giving up their data or their autonomy, and without needing to share confidential identifiers.

For example, imagine that Dr Smith in the USA studies the relationship between P50 brain-evoked potential amplitude and hippocampal volume in a population of chronic schizophrenic outpatients. Smith is not personally interested in genetics, but recognizes the potential value of data sharing, and so collects lymphocytes from each subject and submits these to a DNA repository. Professor Jones in France wants to learn whether polymorphisms in gene X play a role in schizophrenia; so he obtains DNA from this repository and finds that 20% of patients have a particular variant allele as compared to 15% of control subjects—a difference that is not statistically significant. In the present climate, in which tissue banks generally provide only simple information such as age, sex and diagnosis, the analysis might end there with a negative result. However, the linking facility allows Smith and Jones to be aware of each other, and when they perform a pooled analysis of their combined primary data, an entirely different result emerges: Jones finds that hippocampal volume is significantly decreased in schizophrenic individuals with the variant allele.

One of the major advantages of such a facility is that it links different investigators directly, and thus allows independence from the nature or organization of patient data. Although tissue banks have complex data-management systems and databases with idiosyncratic features, the linking facility does not need to be concerned with this level of data organization once researchers have become aware of each other and decide to collaborate. Similarly, if two investigators pool their data for analysis, they are likely to consider only a very small subset of this information, which may be transferred to a simple spreadsheet. To link investigators does not in general require that they automatically link their entire databases. In essence, the linking facility is merely a registry of samples, with each sample from an individual given the same code, together with information on who is using them for what studies.

A potential pitfall of this approach is the possibility that if the same individual enrols in two different research studies, he or she may be assigned two different linking numbers. To minimize this risk, investigators need to routinely ascertain whether subjects have participated in prior studies. In addition, quality control procedures would need to be implemented to detect possible duplications when two investigators pool their data, for example if two different linking numbers share the same name or address. Alternatively, because names and addresses may be written in different ways and may change over time, non-identifying information could also be used, such as date of birth, medication history, ventricular volume or genotype. This issue is a specific example of the record-linkage problem, which has been extensively analysed in other contexts (Winkler, 1995). Nevertheless, the problem is rather minor from the standpoint of analysing pooled data: in the example given above, even if the data from one subject were mistakenly thought to come from two different subjects, this would not alter the inferred relationship between hippocampal volume and allele type.



The ability to link databases on demand would greatly assist collaborative data-mining. The NeuroSys project in the USA (Pittendrigh & Jacobs, 2002) and the Axiope project in Europe (Cannon *et al*., 2002) are both developing tools to help scientists record their experiments directly in a semi-structured format and to publish their data—or at least a description of it—on the web. In the Axiope scheme, each person putting information onto the web retains ownership of their data, but would register at a central catalogue site that can be searched by others looking for studies of a particular type (Cannon *et al*., 2002). Although this web-based approach is elegant, generally applicable and powerful, scientists will need to substantially alter their habits of record-keeping and publishing for peer-to-peer data sharing to succeed. In contrast, the linking facility for banked samples would not require any effort on the part of the scientists involved, other than attaching an additional ID number to the data they already provide. Furthermore, scientists would not need to disclose the nature of their studies beyond the very small community of scientists using the same samples.

At present, many tissue banks are populated by a small number of donating investigators—often only by the group responsible for maintaining the bank. In part, this reflects the specialized nature of certain tissue collections (for instance, tumour tissue removed at the time of surgery) and the need to maintain uniform protocols on tissue handling. However, as it is relatively easy for clinical investigators to obtain and process blood samples, so an enormous increase in the number of banked DNA and plasma samples could be achieved if the following criteria were met: first, repositories would have to be available to handle a large number of samples; second, funding would need to be available to defray the costs of shipping samples to the banks; third, subject consent forms for clinical studies would have to be expanded to allow banking of samples (Malone *et al*., 2002); last, and most important, researchers would have to be convinced that it is to their own direct advantage to allow others to perform studies on their samples. Having a centralized linking facility would particularly help in achieving this last criterion. It would also remove the need to make a single international tissue bank or series of national banks, which would be expensive and cumbersome to operate, and perhaps politically difficult to establish. Instead, the linking facility could coordinate efforts across any number and type of cooperating regional or disease-specific banks, and in this way could truly maximize our potential to understand the underlying causes of multifactorial diseases.

ACKNOWLEDGMENTS

I thank V. Torvik and C. Yu for their comments. N.S. is supported by the National Library of Medicine and the NIH Human Brain Project.

REFERENCES

Cannon, R.C., Howell, F.W., Goddard, N.H. & De Schutter, E. (2002) Non-curated distributed databases for experimental data and models in neuroscience. *Network*, **13**, 415–428.

Geschwind, D.H. *et al*. (2001) The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric conditions. *Am. J. Hum. Genet.*, **69**, 463–466.

Lavori, P.W. *et al*. (2002) Principles, organization and operation of a DNA bank for clinical trials: a Department of Veterans Affairs cooperative study. *Control. Clin. Trials*, **23**, 222–239.

Malone, T., Catalano, P.J., O'Dwyer, P. J. & Giantonio, B. (2002) High rate of consent to bank biologic samples for future research: the Eastern Cooperative Oncology Group experience. *J. Natl. Cancer Inst.*, **94**, 769–771.

Pittendrigh, S. & Jacobs, G. (2002) NeuroSys: a semi-structured laboratory database. *Neuroinformatics* (in the press).

Torrey, E.F., Webster, M., Knable, M., Johnston, N. & Yolken, R.H. (2000) The Stanley Foundation brain collection and Neuropathology Consortium. *Schizophr. Res.*, **44**, 151–155.

Van Horn, J.D. *et al*. (2001) The Functional Magnetic Resonance Imaging Data Center (fMRIDC): the challenges and rewards of large-scale databasing of neuroimaging studies. *Phil. Trans. R.. Soc. Lond. B.*, **356**, 1323–1339.

White, A.W. and Messler, C. (1997) Facilitating linkage through universal patient identifiers: a difficult endeavor. *Top. Health Inf. Manag.*, **17**, 32–39.

Winkler, W.E. in *Business Survey Methods* (eds Cox, B.G. *et al*.) 355–384 (1995) (Wiley Press, New York).

Neil R. Smalheiser is at the Department of Psychiatry MC912, University of Illinois at Chicago, USA.
E-mail: smalheiser@psych.uic.edu

doi:10.1038/sj.embor.embor744