MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

Linking and Analysing Health Data with Appropriate Privacy and Security

Norm Good[a], David Hansen[b], Christine M O'Keefe[c]

[a] *E-Health Research Centre and Mathematical and Information Sciences, CSIRO, Australia*
[b] *E-Health Research Centre, CSIRO, Australia*
[c] *Preventatative Health National Research Flagship and Mathematical and Information Sciences, CSIRO, Australia*

Abstract

In this paper we discuss current approaches for linking and analysing sensitive health data for research, and discuss associated security and privacy risks. We also provide an overview of two CSIRO tools, HDI™ and PPA®, developed for enabling secure access to personal health data for research while providing the required privacy protection, in some situations.

Keywords:

medical record linkage; privacy; statistics; health research

Introduction

The increasing number and size of health data repositories can provide the health and medical research communities with information which is important to clinical treatment and medical research. There is a need not just to access the data, but to locate records relating to the same patient in multiple databases and bring the records together for analysis and research.

Various approaches and protocols have been developed for attempting to simplify the linkage and provision of separate data sources to end users for analysis and research whilst preserving the privacy of individuals, including two and three party protocols and blind record linkage algorithms [1]. The technological approaches differ in their architecture and software/hardware platforms, and there are differences between the applicable legislative and policy frameworks.

Model architectures range from single large databases[2], separate databases linked only as required[3], chains of links type data systems[4], and federated databases[5]. In most cases access to the data is controlled by ethical review boards and the released data is usually in de-identified format, with names, addresses and dates of birth removed.

The role of data custodians in the provision of health data to users varies: with some custodians releasing the data completely and others entrusting third parties to manage the privacy and security issues. As the number of data sources increases, the complexity of privacy and security agreements will obviously increase, placing a substantial amount of responsibility upon the parties involved [6].

Once a linked data set is available, usually in de-identified format, the analysis of the data is generally carried out using statistical or reporting software. In some cases results of subsequent statistical analyses are assessed for disclosure risk prior to publication [2], whilst in others the burden resides upon the trusted third party undertaking the analysis. Whilst ethics agreements may contain clauses governing how results of statistical analyses may be disseminated, there remains a real risk of disclosure [7].

In this paper we propose one solution for linking and analysing health data that may reduce the risk of identifying individual records in some situations.

Health data integration™ (HDI™)

The core functionality of HDI™ enables linking of patient records from different data repositories whilst maintaining the privacy of the patients. It is a virtual repository, and external release of the data is controlled by the data custodian, and can be stopped at any time. HDI™ provides support in software for data custodians to develop a strong governance framework providing rules for: release of data for use in projects; management of the HDI™ installation; support for the separation of identifying and clinical information, and the building of a virtual linked data set by Project Administrators for Users to access linked data. HDI™ offers this data on-line, allowing real time query and retrieval without using a data warehouse.

The HDI™ model allows data custodians to release demographic information to HDI™. As HDI™ is installed locally there is also the provision for third party software to link directly to HDI™ and the data. Within the governance framework provisions can be made that only certain types of analyses may be performed on sensitive data.

Privacy-preserving analytics® (PPA®)

Privacy-Preserving Analytics® (PPA®) is a menu-driven web-based software package for conducting statistical analysis on raw microdata in a secure computing environment. Outputs of an analysis are filtered before

presentation, with the aim that no individual unit record is disclosed or can be inferred.

Reducing the risk of disclosure is achieved in several ways:

- Variable names that are a privacy risk are de-identified
- Some models are fitted on subsamples of the dataset to limit attempts at re-engineering
- Some data transformations are restricted
- Some parameter estimates are rounded, standard errors and exact significance values are not produced, and
- Some higher order interactions are limited.

PPA™ can be used as a knowledge discovery tool for testing hypotheses before developing a full research proposal.

The use of PPA™ may reduce (though will not entirely remove) the administrative burden of consent provisions, custodian security compliance, ethics committee processes and address the requirements of HIPAA (Health insurance Portability and Accountability Act 1996 (US)).

Analysis of health data in an HDI™ and PPA® framework

As an example of HDI™ and PPA® working together, consider that in the HDI™ environment, data custodians may allow HDI™ access to sensitive information such as individual demographic details, surgeon, and hospital. In that case, a flag may be attached to a sensitive field to inform HDI™ that this field may be used as an explanatory variable in an analysis using PPA® but no estimate may be published. PPA® can then use that information for constructing and running a full statistical model to gain reliable information about other explanatory variables, where the estimates can be published. In a knowledge discovery context a researcher can test hypotheses regarding the effect of factors such as age, gender and comorbidities on clinical outcomes, for example, while having the effect of the surgeon modeled but not released.

Conclusion

A data custodian has responsibility to ensure that no identifiable information is released to an unauthorised third party. Project agreements normally require that data is used only for the specific aims of the project for which it was released, and that the data is destroyed at the end of the project. The release of results from the project must also satisfy privacy agreements. This is the limit of control a data custodian has over their data once it has been released. The HDI™/PPA™ model may support the fulfillment of a significant amount of the responsibility of the data custodian by significantly reducing the risk that identifiable information is publicly released from a project. Similarly, researchers may benefit by being able to use more sensitive and valuable information for testing hypotheses, though without actually seeing the data. The HDI™ and PPA™ tools may also allow less authorised third-parties such as graduate students, and other researchers to conduct analysis on raw data without having to modify existing data agreements with data custodians.

Acknowledgments

HDI™ and PPA® are projects of the CSIRO Preventative Health National Research Flagship. HDI™ is being developed by researchers at the e-Health Research Centre, a collaboration between CSIRO and the Queensland Government and PPA® is being developed by researchers in CSIRO Mathematical and Information Sciences.

References

[1] Agrawal, R., Evfimievski, A., and Srikant, R. (2003). Information sharing across private databases. In SIGMOD 2003, 86-97. ACM.

[2] Manitoba Centre for Health Policy: Population Health Research Data Repository. http://www.umanitoba.ca/centres/mchp/protocol_external/databases.shtml [accessed 7 July 2007]

[3] University of British Columbia: BC Linked Health Database. http://www.chspr.ubc.ca/data [accessed 7 July 2007]

[4] University of Western Australia: WA Data Linkage Unit. http://www.publichealth.uwa.edu.au/welcome/research/dlu/linkage [accessed 7 July 2007]

[5] Bio21:MMIM: Molecular Medicine Informatics Model. http://mmim.ssg.org.au/whatis.htm [accessed 7 July 2007]

[6] Patilla, J.C. (2001). Security must be baked in, not painted on. eEWeek.com, 6 August. http://www.eweek.com/article2/0, 1759,465346,00.asp [accessed 29 Nov 2006]

[7] Mole, D., and Fox, C. (2005). Electronic data protection: Procedures need drastic improvement. British Medical Journal. 330, 537.