HIC 2008 Australia's Health Informatics Conference
Editor: Heather Grain
Publisher: Health Informatics Society of Australia Ltd (HISA)
ISBN 978 0 9805520 0 3

Secure Data Linkage and Information Sharing With GRHANITE™

Siaw-Teng Liaw PhD, FRACGP, FACHI[1], Douglas Boyle PhD, FACHI[2]

[1]Professor, [2]Senior Research Fellow, School of Rural Health
University of Melbourne Faculty of Medicine, Dentistry & Health Sciences.

Abstract

Objectives:

To address and simplify the operational, ethics, consent and governance processes during the secure collection of individual identifiers and the accurate, reliable and confidential identification and re-identification of individuals to facilitate safety and quality of health care as well as accurate and rigorous evaluation research.

Background:

Linking accurate personal information to decision support systems can improve professional services e.g. health care or financial management. Linking professions will promote integrated services, avoiding costly duplication of infrastructure and services. However, where person-specific information is held, there are risks of identity theft or confidentiality breaches. The ethics and governance processes involved in routinely collecting data for service provision, research, evaluation, quality assurance, policy development and governance reporting need to be integrated yet simplified.

Methods:

Functional specifications for ethical, secure and accurate information sharing and management were established. A flexible decentralised design methodology was adopted to develop GRHANITE™ to manage informed consent, encrypt, extract, link and manipulate personal clinical data without collecting or exposing personal identifiers.

Results:

GRHANITE™ reduces risks of confidentiality breaches and identity theft by dispensing with the need to collect and store recognisable person identifiers and by forcing informed consent processes. GRHANITE™ has interfaced with disparate technologies in a generic manner to: (1) demonstrate the secure, de-identified linkage processes enable accurate and reliable identification; (2) de-identification mechanisms with security protocols can effectively guarantee privacy in the collection, linkage, aggregation and analysis system; and (3) Secure re-identification of individuals is still possible in source systems. Large-scale de-identified data repositories which hold no identifiable person-identifiers BUT are able to perform automated data linkage and acquisition can be developed and maintained in a viable manner.

Implications:

The GRHANITE™ research and development program will provide the knowledge to underpin the next generation of encrypted data repositories and their implementation in service organisations, delivering innovations in personal identification, information protection and integrity of national security systems.

HIC 2008 Australia's Health Informatics Conference
Editor: Heather Grain
Publisher: Health Informatics Society of Australia Ltd (HISA)
ISBN 978 0 9805520 0 3

Objectives/hypotheses:

To describe the principles, methods and tools developed to simplify the operational, ethics and governance processes, and standard operating protocols during the secure and confidential collection of individual identifiers, accurate and reliable identification and re-identification of individuals. A secondary objective was to demonstrate that large-scale de-identified data repositories with the capability to perform automated and generic data linkage and acquisition can be developed in a viable manner.

Background/literature review:

According to the "Building and transforming Australian Industries through technical innovation and safeguarding Australia" national research priority, "improved data management is central to the future competitiveness of key industries" and "personal identification, information protection and the integrity of security systems are fundamental towards ensuring the national security of Australia". As part of eCommerce, the use of information and communication technology (ICT) to transmit, store and retrieve digital data electronically for commercial and administrative purposes, locally and at a distance, Government and business routinely access data repositories for information to support routine transactions, quality assurance, strategic & financial planning and reporting for governance and regulatory purposes. Vast amounts of social and personal information, including tissue samples, about individuals and populations are being collected AND innovative and sophisticated ways of linking and manipulating the information in these repositories are being developed. Where person-specific information is held, there is a risk of identity theft or other abuses of information. A greater understanding of the design and management of large data repositories is required to achieve the above national priorities with minimal risk to citizens and public systems. Data protection breaches can be due to errors of omission, e.g. when the NSW Transit Authority sold its old computers without wiping the hard disks in 2005, or commission, e.g. when the US Veterans Affairs data base protection was intentionally breached by a malicious employee in 2006. Deliberate network attacks have received regular public airings as had the recent loss of personal (and banking) details regarding 25 million individuals in the UK due to 2 compact disks going astray[1]. The theft of laptop computers holding sensitive information, e.g. theft of a notebook computer belonging to Princess Diana's psychiatrist, are unfortunate but relatively frequent occurrences (3.4% per annum) according to NSW Bureau of Crime Statistics[2].

This research uses decentralized security-focused software design, encryption techniques and good standard operating procedures to minimize the risks and impact of confidentiality breaches when these incidents occur. The efficient, safe and ethical development and use of data repositories need to consider:

- The consent of individuals is usually required and the variations in extent and nature of consent, including any changes with time, must be addressed appropriately[3, 4];
- Data linkage allowing information from disparate sources to be attributed to a common individual, requires accurate and reliable identification and authentication of individuals[5], meaning that individual identifiers must be collected at some point;
- Sourcing data from disparate sources implies a capability to interface with many disparate technologies in a very generic manner;
- Large-scale linkage and interfacing implies that communications must be automated and often require external connections, usually through corporate firewalls. Security must be inherent in the collection, linkage, aggregation and analysis system; and
- After data aggregation and analysis, it is often necessary to contact individuals for specific reasons, e.g. the follow-up of financial transactions or an abnormal health event. Re-identification is often required.

Data extraction from information repositories, record linkage, identification, encryption and decryption, and analysis within an ethical and flexible consent framework is a central component of our clinical, community health and informatics research program (www.conduit.unimelb.edu.au), which aims to improve the quality and utility of routinely collected data within a secure and private environment. This "cleaned and linked"

ISBN 978 0 9805520 0 3

information will be used for clinical information-sharing, computerized decision support[6], clinical audits and quality control, data mining, longitudinal population-based research[7] and knowledge interchange[8]. Evidence-based decision-making to facilitate safety and quality of care will be more timely and cost-efficient. Information linkage among the professions and disciplines in all sectors of society will promote greater integration of services leading to greater cost-efficiencies from less duplication of infrastructure and services, more accurate and rigorous evaluation research to support quality assurance, risk management, health outcomes monitoring, policy development and service planning.

In the real world of disparate legacy systems and lack of implemented interoperability standards, pragmatic interfacing and aggregating mechanisms that can circumvent the lack of standardisation and provide an affordable migration path for data from legacy systems into newer technologies as they become available, are essential.

Methods:

The GRHANITE™ (Generic Health Network IT in the Enterprise) approach recognises that while interoperability (technical, syntactic, semantic and business) standards are desirable for modern interfacing applications, there are a large number of legacy systems in existence that cannot conform to these standards for technical or financial reasons. The GRHANITE™ suite of utilities was systematically developed to provide a generic data extraction and record linkage capability, within a confidential, secure and ethical framework. Using Microsoft C# and SQL Server 2005 as the software platform and the latest validated industry standards for encryption[9-11], GRHANITE™ brings together a range of publicly available tools and techniques to manage informed consent processes and the linkage of records in large databases safely and securely. Novel combinations of technologies were developed for consent management, interfacing, de-identification and record linkage. The emphasis is flexibility to deal with issues such as database technology diversity, security restrictions, firewalls and IP address translation. RSA[11] and AES[10] encryption and key exchange was used to allow secure TCP/IP communication between a central GRHANITE™ Authority Service and interfaces and tools that run and communicate to databases or flat files using ODBC[12], OLEDB[13], SQL Server, Oracle and text parsing technologies. A Service-Oriented Architecture (SOA) was used[14, 15] because it allowed GRHANITE™ to circumvent most firewall issues as most organisations will permit web access. A flexible decentralised methodology was adopted to ensure the rights and confidentiality of the individual client and provider when collecting, linking and analysing their information currently available in administrative, financial and health information systems.

Results:

GRHANITE™ has been developed to a fully functional prototype, with these key features:

1. Modular and distributed within a web service architecture
The GRHANITE™ Authority Service (1) communicates with a central web service that brokers the uploading (and downloading) of data; (2) checks the consent status of all records being accessed by the GRHANITE™ interfaces; and (3) deals with a variety of databases by using encrypted XML files definitions that stipulate what, when and which database can be connected to, and what fields and files are to be extracted. The ability to send data to different databanks means that data to be used for different purposes does not need to reside in one large repository.

2. Secure information storage
All patient-identifying data in GRHANITE™ are encrypted at all times, with users responsible for defining data extractions or consenting individuals needing to log-in to set-up or perform any function. Audit trails and anti-tamper techniques are used to protect the system against unauthorised database updates and hacking. No identifiers ever leave clinic / practice computer systems.

ISBN 978 0 9805520 0 3

3. Secure communications

Because the GRHANITE™ Authority and client PC initiate all communications within the web services architecture, the security of the organisation is not compromised. In addition to standard SSL web service encryption[16], GRHANITE™ uses cycling key generation and other techniques as a part of every message. The web server is incapable of data decryption so even if the web server is compromised, the data arriving at the web server cannot be compromised. As a further safeguard, the data can be electronically forwarded or manually transferred to secure databanks for decryption, ensuring electronic isolation between the web server and the secure databank. GRHANITE™ databank automatically create tables and perform record linkage without any human intervention, reducing the likelihood of human security breaches.

4. Ethical record linkage

GRHANITE™ supports a range of consent models - no consent (a waiver of consent), opt-in consent or opt-out consent - according to different requirements. GRHANITE™ also utilises information from consent fields of existing electronic databases. It manages consent across an enterprise so that a client who denies consent will stop the transmission of data from any computer in the network and ensure any existing data is removed from central databanks in the enterprise.

The inherent security of GRHANITE™, along with relevant Standard Operating Procedures governing data collection, storage and use, has enabled the approval of "opt-out consent" for the CONDUIT record linkage research program; the "opt-out" consent methodology is currently being evaluated in a number of health services.

5. Secure record linkage

GRHANITE™ automatically generates Statistical Linkage Keys (SLKs) at the time of data extraction with no person identifiers ever leaving the source databases. A combination of advanced encryption technologies generate these non-reversible SLKs, based on combinations of person identifiers such as surname and date-of-birth[17]. Techniques have been implemented to ensure non-registered computers and sites are not able to generate these keys. GRHANITE™ uses a variety of proprietary techniques to improve the sensitivity and specificity of record linkage far beyond that traditionally found in hashed deterministic linkage. The efficacy of the GRHANITE™ techniques have been statistically ascertained through a comparison with an industry standard probabilistic record linkage technology (from Sun Microsystems) linking 45,000 records from 10 different data sources. An extensive in-house test, using a database of 850,000 individuals in Victoria and the 2006 Australian Census[18] have further confirmed the efficacy and accuracy of the GRHANITE™ SLK methodology. The GRHANITE™ record linkage methodology uses some techniques related to those described by Agrawal et al [28], Churches & Christen [19] and O'Keefe et al (2004) [20]. Details of the specific techniques employed will be reported in an up-coming paper. GRHANITE™ uses the security inherent in its implementation to achieve a middle ground between absolute anonymity and pragmatic de-identified record linkage.

6. Tested technology and peer-reviewed methodology

GRHANITE™ is currently being piloted in a national Chlamydia Surveillance project in 68 sites (30 general practices, 30 laboratories and 8 family planning centres) across Australia (www.burnet.edu.au/home/cephr/current/access). The security focused design of GRHANITE™ was a factor in this project being approved for a waiver of consent from the Royal Australian College of General Practitioners National Research and Evaluation Ethics Committee. This highlights the fundamental importance and practical advantages that can be facilitated by information security and ethics technologies like GRHANITE™.

In summary, GRHANITE™ has been interfaced with disparate technologies and systems in a generic manner to: (1) demonstrate the secure collection of individual identifiers to enable the accurate and reliable identification of an individual, (2) embed security in the collection, linkage, aggregation, and analysis tools, and (3) enable the secure re-identification of individuals for safety and quality reasons in source systems.

HIC 2008 Australia's Health Informatics Conference
Editor: Heather Grain
Publisher: Health Informatics Society of Australia Ltd (HISA)
ISBN 978 0 9805520 0 3

Discussion of Implications:

GRHANITE™ is a general and comprehensive data interfacing application to manage data across an enterprise BUT its uniqueness and significance is that it dispenses with collecting and storing person identifiers! The automatic generation of 'keys' for record linkage and storage at the time of data extraction has dispensed with the need for a third party key-holder and the complex governance structures associated with key-holder mechanisms. This greatly simplifies the practical aspects of maintaining centralised data repositories. For instance, if the United Kingdom had utilised GRHANITE™ technologies, there would be no need to store NHS numbers, names and other identifiers on the national 'spine' data repository. Encrypted databases that do not collect and store person identifiers will enhance the security and confidentiality of the proposed Australian shared electronic health record. The risks and consequences of wide-scale identity theft and confidentiality breaches will be greatly minimised.

The GRHANITE™ informed consent management tool, encryption techniques and explicit and transparent standard operating procedures provide a strong argument to support the use of "opt-out" consent models for information sharing for clinical and research purposes. This approach builds on the Scottish Care Information – Diabetes Collaboration (SCI-DC) (www.crag.scot.nhs.uk/topics/diabetes/diabit/q&a.doc), which has addressed the ethical and interfacing barriers to information exchange across all care providers for all patients with diabetes in Scotland. It is worth noting at this point that the lack of international and national standards for data linkage and database interoperability, along with variable governance and ethical processes, will inevitably lead to incomplete and inaccurate clinical and research information, leading to poor continuity and quality of care as well as inaccurate research.

There is a research translation and knowledge transfer imperative to gather the evidence to support the efficacy and effectiveness of technologies like GRHANITE™, and make these tools available to all organisations in an easy to use-and-adopt software suite. Decentralisation of "keys" management to source organizations is both empowering and cost-efficient, encouraging individuals and organizations to promote and sustain information security and privacy in their practice. It can be argued strongly that all linked and centralised data repositories should use these technologies.

The GRHANITE™ program also increase the informatics competencies of professionals, researchers, teachers and students, contributing to Australia's capacity to successfully adopt the eCommerce, eHealth, eResearch and eLearning applications and approaches required to become a successful knowledge nation generally.

Further research is required to formally model the GRHANITE™ system security, internal and external breaches, audit procedures and processes to mitigate risks, and the impact of incorporating hashing techniques to de-identify person identifiers, HL7 messages[21], and to prove linkage specificity and sensitivity. Further development will include consideration of automated dataset extraction with disclosure control techniques such as k-anonymity and variations such as l-diversity or t-closeness[22-24], use of Security Enhanced (SE) Linux to develop isolated Ultra-Secure repositories[25], consumer digital certification and trustee infrastructures[26], and utilities such as Shibboleth web single sign-on combined with PERMIS role management[27].

The GRHANITE™ research program will generate knowledge to underpin the next generation of encrypted repositories and implementations with enhanced security, confidentiality, record linkage and generic interfacing capabilities. This will improve data management in service organisations and deliver innovations in personal identification, information protection and integrity of national security systems as well as simplify the governance of information sharing within and among organisations. The improved privacy and security arrangements during information sharing will encourage acceptance of information sharing by consumers and professionals for clinical, health care, audit, policy, research and community health purposes.

HIC 2008 Australia's Health Informatics Conference
Editor: Heather Grain
Publisher: Health Informatics Society of Australia Ltd (HISA)
ISBN 978 0 9805520 0 3

References:

[1] Anonymous. Our information was put at risk. 2007 [cited 2008 2008 Feb 28]; Wednesday 21st Nov 2007:[Available from: http://news.bbc.co.uk/1/hi/uk/7106336.stm
[2] Analytics C. Calson Analytics note consumer data losses. 2008 [cited 25/02/2008]; Available from: http://www.caslon.com.au/datalossnote3.htm
[3] Coiera E, Clarke R. e-Consent: The design and implementation of consumer consent mechanisms in an electronic environment. J Am Med Inform Assoc. 2004 2004 Mar-Apr;11(2):129-40.
[4] Liaw S, Boyle D. Lessons from the NHS National Programme for IT. MJA. 2007;186(11):607.
[5] Brenner H, Schmidtmann I. Effects of record linkage errors on disease registration. Methods Inf Med. 1998;37:69-74.
[6] National Electronic Decision Support Taskforce. Report to Health Ministers: Electronic Decision support in Australia. Canberra: National Health Information Management Advisory Committee; 2002 November 2002.
[7] van Weel C. Longitudinal research and data collection in primary care. Ann Fam Med. 2005 2005 May-Jun;3(Suppl 1):S46-51.
[8] Liaw S, Schattner P. eConsulting. In: 141. TREV, ed. Methods in Molecular Medicine series: Clinical Bioinformatics. Totowa NJ: Humana Press, 2008:353-73.
[9] Anonymous. Secure Hash Algorithms (SHA) hash functions. 2008 [cited 28 Feb 2008]; Available from: http://en.wikipedia.org/wiki/SHA_hash_functions
[10] Federal Information Processing Standards Publication. The Advanced Encryption Standard (AES); 2001.
[11] Rivest R, Shamir A, Adleman L. A Method for Obtaining Digital Signatures and Public-Key Cryptosystems. Communications of the ACM. 1978;21(2):120-6.
[12] Open Database Connectivity (ODBC). Open Database Connectivity. 2008 [cited 25/02/2008]; Available from: http://en.wikipedia.org/wiki/Open_Database_Connectivity
[13] Anonymous. OLE DB Object Linking and Embedding, Database 2008 [cited 28 Feb 2008]; Available from: http://en.wikipedia.org/wiki/OLE_DB
[14] Papazoglou M, van den Heuvel W-J. Service oriented architectures: approaches, technologies and research issues. The VLDB Journal. 2007(16):389–415.
[15] Yunus M, Mallal R. An Empirical Study of Security Threats and Countermeasures in Web Services-Based Services Oriented Architectures. In: Kitsuregawa M, ed. Web Information Systems Engineering - WISE 2005. Berlin/Heidelberg Springer 2005:653 – 9.
[16] Anonymous. Transport Layer Security. 2008 [cited 25/02/2008]; Available from: http://en.wikipedia.org/wiki/Secure_Sockets_Layer
[17] Gilbert H, Handschuh H. Security Analysis of SHA-256 and Sisters. In: Matsui M, Zuccherato R, eds. Selected Areas in Cryptography. Berlin / Heidelberg: Springer-Verlag 2004:175-93.
[18] Australian Bureau of Statistics. 2006 Census. 2008 [cited 28 Feb 2008]; Available from: www.abs.gov.au
[19] Churches T, Christen P. Some methods for blindfolded record linkage. BMC Med Inform Decis Mak. 2004 2004 Jun 28(4):9.
[20] O'Keefe C, Yung M, Gu L, Baxter R. Privacy-preserving Data Linkage Protocols. Workshop on Privacy in the Electronic Society '04; 2004 October 28, 2004; Washington, DC, USA; 2004.
[21] HL7 Australia. Health Level 7 Australia. 2008 [cited 28 Feb 2008]; Available from: www.hl7.org.au
[22] Li J, Wang H, Jin H, Yong J. Current Developments of k-Anonymous Data Releasing. The National e-Health Privacy and Security Symposium 2006 (ehPASS'06); 2006; 2006.
[23] Li N, Li T, Venkatasubramanian S. t-Closeness: Privacy beyond k-Anonymity and l-Diversity. IEEE 23rd International Conference on Data Engineering; 2007 April 2007; 2007. p. 106-15.
[24] Machanavajjhala A, Gehrke J, CKifer D. l-diversity: privacy beyond k-anonymity. . The 22nd International Conference on Data Engineering (ICDE06); 2006; 2006. p. 24.
[25] Henricksen M, Caelli W, Croll P. Securing Grid Data Using Mandatory Access Controls. ACSW '07: Fifth Australasian Symposium on Grid Computing and e-Research (AusGrid 2007); 2007; 2007. p. 25 - 32.
[26] Au R, Croll P. Consumer-Centric and Privacy-preserving Identity Management for Distributed e-Health Systems. 41st Hawaii International Conference on System Sciences; 2008; Hawaai; 2008.

HIC 2008 Australia's Health Informatics Conference
Editor: Heather Grain
Publisher: Health Informatics Society of Australia Ltd (HISA)
ISBN 978 0 9805520 0 3

[27] Jie W, Huang Z, Daw M, Procter R, Li X, Tang L, et al. Secure Access to Grid Information Service Using Shibboleth and PERMIS. The 9th IEEE International Conference on E-Commerce Technology and The 4th IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services (CEC-EEE 2007); 2007 23-26 July 2007; 2007. p. 297 - 304.