Lumos Technical Information Summary

Appendix to Lumos GP Information Pack

Version 2.0 - February 2020

Introduction

Lumos is a pioneering initiative that links data from general practice to hospital, community, mortality, and other NSW data collections. It is a large-scale, transformational program that sheds light on the patient journey through the NSW health system; marking the first of its kind, at scale, in Australia. Lumos data linkages will be carried out using ‘privacy preserving record linkage’ (PPRL) whereby patient names and other personally identifying particulars are fully encoded before extraction from the source practices. The encoding used in PPRL is one-way, and names cannot be reconstructed from the coded information thereby de-identifying the linkage process. Lumos will still link records probabilistically (i.e. allow for slight differences in patient details) and apply the standard ‘principle of separation’ whereby the information used to link records is separated from the health content information at all stages of the process. The other major innovation of Lumos is the development of a secure platform for data storage and access of the resulting linked data to ensure that data is stored and accessed with the highest level of security available.

This appendix outlines the key components of the solution in the context of General Practice.
Participating General Practices will receive personalised reports, specifically for their practice generated by the NSW Ministry of Health and distributed by PHNs after each bi-annual linkage.

General Practice Data

Several measures are in place to protect general practice records. In a general order of sequence:

1. Informed consent to participate in the ethically approved linkage is required from the general practice data custodian(s) as the very first step.
2. When an extract is triggered for linkage, the ‘data extraction tool’[1] performs the following steps **within** the general practice:
    a. The data is extracted from a compatible ‘clinical management system’[2]
    b. The patient identifying particulars then undergo a complex encoding process, in which a one way cryptographic hash function obfuscates identifiers while allowing for probabilistic linkage – this process of cryptographically ‘encoding at source’ is fundamental to the ‘privacy preserving record linkage’. The encoding algorithm used for this purpose has been developed by Curtin University (used under a commercial license), and is described with selected excerpts below from JH Boyd, "*Record Linkage Techniques: Exploring and developing data matching methods to create national record linkage infrastructure to support population level research*"[3]:
        i. Privacy preserving record linkage using Bloom filters works by encoding personally identifying information into Bloom filters (binary vectors). A Bloom filter begins as an array of a set length, with all elements set to zero. Each field (e.g. first name) is broken down into overlapping sets of letters (qgrams). Padding is often used to give the first and last letters their own bigrams. Each of these qgrams is passed through a series of cryptographic hash functions. A hash function is an algorithm which produces a fixed-length output with several important properties. Firstly, given the same input, it will always produce the same output (i.e. the same qgram will produce the same hash value). Secondly, the hash function is one way, meaning it is not possible to determine the encoded qgram from any given hash value (i.e. it is irreversible). Different hash passwords can be used to produce different output. The modulus of these hashes is then computed with respect to the length of the Bloom filter. This process allows us to map each bigram to a position in the Bloom filter. These positions are then set to 1. Two Bloom filters can be compared to each other using a dice coefficient. The Sørensen–Dice coefficient is calculated as twice the number of positions in which both Bloom filters have a value of one, divided by the number of positions set to 1 in total. The dice coefficient results in a score between 0 and 1, where a higher score reflects greater similarity. The encryption techniques used in privacy preserving linkage with Bloom filters means that probabilistic type techniques can be used during the matching process.
        These techniques allow for small errors such as spelling mistakes which greatly improve linkage quality. Evaluations using real data found linkage quality using Bloom filters to be equivalent to those achieved using unencrypted personal identifiers, and greater than that of other implemented privacy preserving methods.

[1] As of publication, PenCS PenCAT is the only compatible data extraction tool.
[2] PenCAT currently supports extraction from Best Practice, Medical Director and Zedmed for Lumos.
[3] https://espace.curtin.edu.au/handle/20.500.11937/54163

   c. Once the required fields have been encoded using the process above, and other applicable fields have been encoded using standard secure hash algorithm (SHA), the records are split into two different files to adhere to the separation principle. This "principle of separation" is another key mechanism to support best practice privacy preserving record linkage. One file contains the encoded identifiers and the other file contains the 'content data' i.e. visit dates, MBS codes, selected diagnoses, blood pressure, HbA1c values, height, weight, etc. Each of the two files contains a computer generated ID field which allows the data to be securely re-linked at the appropriate stage (detailed further on). This means the content data never travels with the encoded identifiers, which further preserves privacy.
   d. To transfer data securely, the CHeReL utilises a secure File Transfer Protocol (FTPS) with support for Transport Layer Security (TLS). TLS is commonly used by websites for Hypertext Transfer Protocol Secure (HTTPS). Only current TLS versions are used. It leverages Implicit FTP which allows it to be easily accessed without changing firewalls or reconfiguring networks. A certificate and private key (used to perform the 'handshake' using a client x509 certificate), is required to establish a secure connection.
      Authentigate is a certificate exchange portal that has been developed to support the scale-up of Lumos and has undergone an independent privacy and security assessment. It implements the following process:
      i. Verify the general practice is ready to be registered and the server knows who the practice is without giving up the identity of the general practice. This is achieved by hashing the custodian ID using SHA512.
      ii. Obtain a valid certificate by sending a Certificate Signing Request to Authentigate for signing. A valid certificate will be returned in Privacy Enhanced Mail (PEM) format. The private Rivest–Shamir–Adleman (RSA) cryptosystem key remains secure at all times and is a minimum length of 2048 bytes (which is industry best practice for key exchange) and can use RSA based SHA256 or SHA512 signing algorithm in generation. The signing request is in Public Key Cryptography Standards #10 (PKCS10) format and encoded using PEM. The returned certificate will have a 2 year expiry. Once the certificate has expired, a process is triggered to recertify.
   e. Each file is a text file encoded using Unicode Transformation Format (UTF-8) containing comma separated values (CSV). Each file is then compressed prior to transmission using Zip archive along with the project definition file and summary source data analysis.

3. The files are sent to the CHeReL. The CHeReL then makes a specific Project Person Number (PPN) for each individual. A different PPN is created for each project, and for each Lumos linkage. This means no records can be linked retrospectively between linkages, which further protects privacy.
   a. This approach to preserving privacy and data governance has been strongly supported by organisations that are custodians of health records, human research ethics organisations, data users and the community. By providing a mechanism to create linked data with encoded personally identifying particulars, the CHeReL enables ethically approved data linkage in the public interest to be carried out without consent, minimising bias and allowing access to de-identified linked data on whole populations.
4. Linked Data is then securely transferred to the System Information and Analytics branch (SIA) within the NSW Ministry of Health (MoH) to manage the process of data pre-processing, cleaning and validation.
5. Storage of the Linked Data – The Secure Analytics Primary Health Environment (SAPHE) has been developed specifically to house the Lumos data.
   a. An independent risk assessment completed by an external consulting agency identified the need for a secure centralised repository to store the linked data assets in order to maintain the integrity of the data and improve standardisation and quality. Archived versions of the datasets will be retained by the MoH for a 3 year period. Upon completion of pre-processing of the linked data sets received from the CHeReL, the linked data asset will be securely transferred to and stored on the SAPHE.
   b. The SAPHE is a custom-built, secure, remote-access computing environment and cloud repository that provides highly secure storage and analytic tools to access the Lumos linked data asset for approved analyses. The SAPHE does not hold any personal data that directly identifies individuals within the Lumos linked data asset and implements controls that restrict access and outputs to further ensure privacy.

c. The SAPHE will be developed in line with the eHealth NSW's stringent Privacy and Security Assurance Framework (PSAF) to ensure it meets the health system's security and functional requirements for storage of sensitive data. Access will be strictly controlled by the SAPHE administration team and overseen by the Lumos Data Governance Committee.

d. The SAPHE architecture incorporates encryption at rest and uses symmetric encryption to encrypt and decrypt large amounts of data quickly. All Lumos data is encrypted at-rest and in-flight.

A diagram summarising this process is provided at the end of this document.

General Practice Variables:

Variables extracted from general practice are strictly in accordance with ethical approval. Ethical amendments may occur over time to include additional variables. General practice variables currently include:
General Practice ID, Record ID, Patient Status (Active, Archived, Deceased), Age, Gender, Clinical Software, Post Code & Ethnicity (to adjust for demographic characteristics), Aboriginality, DVA Status, Marital Status,
Chronic Disease Flags and diagnosis dates (including: Diabetes, Heart Failure, Cardiovascular Disease, Asthma, Chronic Obstructive Pulmonary Disease, Chronic Kidney Disease, Mental Health – ADHD, Autism, Bipolar, Schizophrenia, Depression, Chronic Osteoporosis, Myocardial Infarction, Acute Coronary Syndrome, Stroke, Carotid Stenosis, Atrial Fibrillation, Cancer, Hyperlipidemia, Coeliac Disease, Liver Disease), Medication Flags and dates of last prescription (including: ACE inhibitor, ARB, anti-thrombotics, Beta Blockers, Respiratory, Antidiabetics, Calcium Antagonist, Diuretics, Lipid Modifier, Medication for Osteoporosis, Medication for Mental Health, Antipsychotics, Anticoagulant, NSAIDS, Glucocorticoids, Mineralocorticoids), Medication List with Dates and number of repeats of last prescription, Encounter Dates and Provider Category, MBS Items and Dates, Health risks (Smoking, Alcohol), Measurements and dates (Blood Pressure, Spirometry, Body Mass Index, Height, Weight), Pathology results and dates (including: HbA1C, Cholesterol, HDL, LDL, EGFR, Microalbumin, Triglycerides, Haemoglobin, ALT), MyHR status and Shared Health Summary dates, Immunisation (Influenza, Pneumococcal).

Ethics

In addition to the technical mechanisms to preserve privacy, Lumos operates under ethical approval which further reinforces adherence to privacy standards. The ethics approval reference is detailed in the Lumos GP Consent Form. For further information contact lumos@health.nsw.gov.au

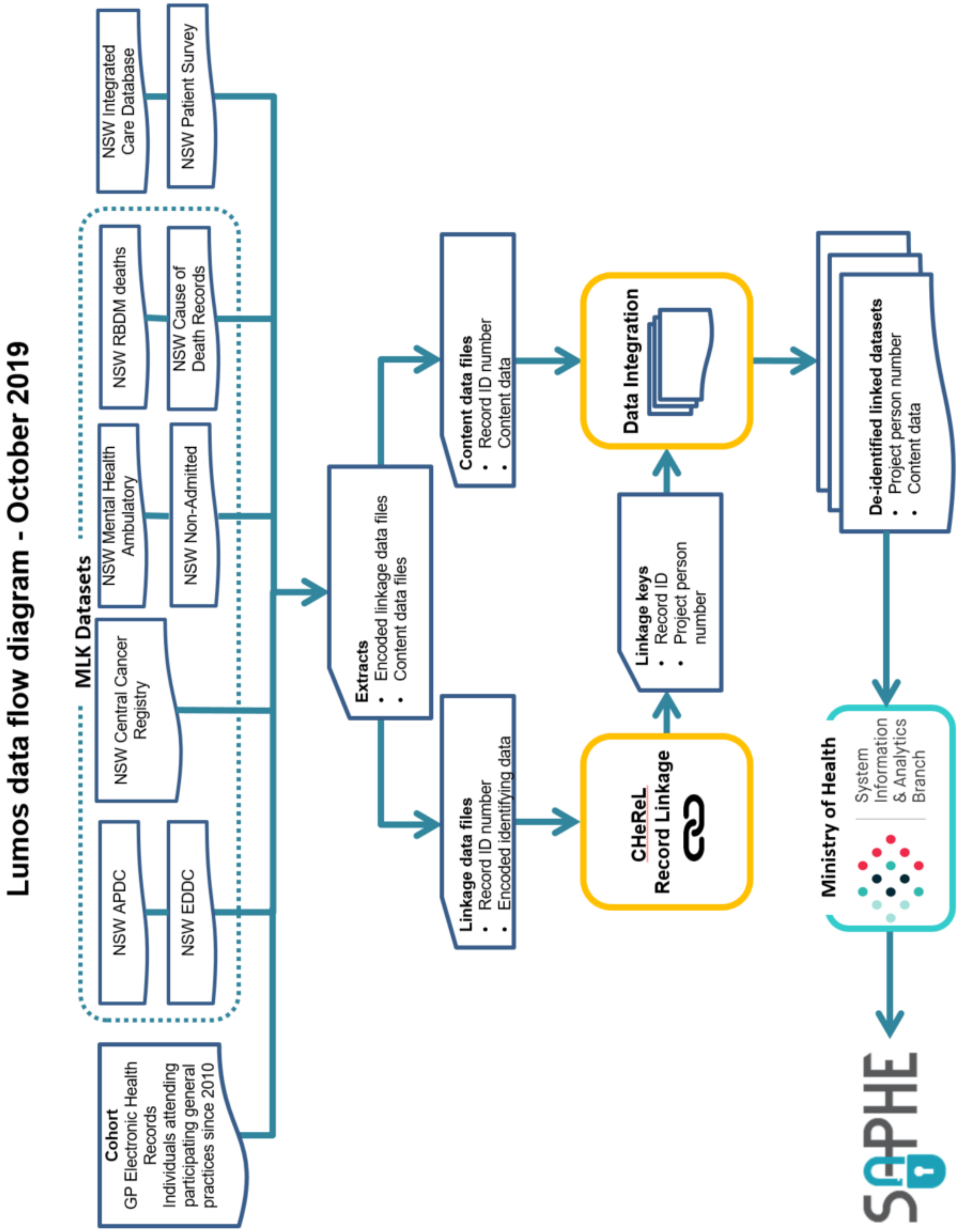
Governance

In addition to the technical and ethical mechanisms to preserve privacy, Lumos operates within a strict data governance framework (DGF) which complies with the following legislation, and a summary of this DGF is presented in the GP Information Pack:

1. Crimes Act 1900 (Cth)
2. Government Information (Public Access) Act 2009
3. Health Administration Act 1982 (NSW)
4. Health Records and Information Privacy Act 2002 (NSW)
5. Higher Education Standards Framework (Threshold Standards) 2015 (Cth)
6. Higher Education Support Act 2003 (Cth)
7. Mental Health Act 2007 (NSW)
8. Privacy Act 1988 (Cth)
9. Privacy and Personal Information Protection Act 1998 (NSW)
10. Public Health Act 2010 (NSW)
11. State Records Act 1998 (NSW)

Further Information

For further information about Lumos, please contact: lumos@health.nsw.gov.au


Lumos data flow diagram - October 2019
Cohort
GP Electronic Health Records
Individuals attending participating general practices since 2010
MLK Datasets
NSW APDC
NSW EDDC
NSW Central Cancer Registry
NSW Mental Health Ambulatory
NSW Non-Admitted
NSW RBDM deaths
NSW Cause of Death Records
NSW Integrated Care Database
NSW Patient Survey
Extracts
• Encoded linkage data files
• Content data files
Linkage data files
• Record ID number
• Encoded identifying data
Content data files
• Record ID number
• Content data
CHeReL
Record Linkage
Linkage keys
• Record ID
• Project person number
Data Integration
De-identified linked datasets
• Project person number
• Content data
Ministry of Health
System Information & Analytics Branch
SAPHE