Public Health e-Labs: A federated model for e-Epidemiology

Citation for published version (APA):

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester’s Takedown Procedures [http://man.ac.uk/04Y6Bo] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.
The problem we address is the general lack of effective and efficient integration of health-related data, methods and expertise, across defined populations, for public benefit. This is a problem for most health systems throughout the world.

Population: We focused on the Northwest of England, which has a rich diversity of social structures, health outcomes and research-active healthcare organisations. In particular, we studied the Salford health economy (population 216,000) because it has relatively advanced healthcare information, with an integrated electronic health record (EHR) across primary care, secondary care and the public sector commissioner/payer.

Objectives and Methods: The requirements for epidemiological integration and processing of data were identified through an iterative process, over two years, involving pilot studies, using crude or enhanced EHR data. Salford clinicians were engaged in each study, an Informatician (Baker) was embedded in Salford, and statistical, epidemiological and software engineering input was provided by the University of Manchester. The criteria that we used for identifying, and engineering, requirements were: governance; patient/citizen privacy; ethics; data quality; skill-gaps; clinical information utility; public health information utility; and scalability.

Findings: A key finding was that the EHR data alone were insufficient for scientific purposes, analyses were enriched by extracting tacit knowledge, through playback of emerging findings to local clinicians, encoding this as explicit metadata, and using the metadata in data-cleansing or statistical-modelling. We found that large-scale record linkage across the health economy was acceptable, provided it took place within the healthcare setting, specifically, as close to primary care physicians as possible. In the UK, ‘Primary Care Trusts’ bring primary care providers and commissioners/purchasers together. Academics worked effectively on the data, in the Primary Care Trust setting, under honorary contracts covering information governance. Data cleansing and organisation was more computationally expensive than anticipated, typically involving re-coding, filtering and re-organising up to 1,000,000 messages. We developed a federated model, using a ‘research object’ paradigm, to provide a scalable architecture for e-Labs.

Conclusion and Recommendations: The outputs of the pilot e-Lab projects were of high value to both healthcare and academic organisations. The Northwest Region of England is now preparing to form a confederation of e-Labs, known as ‘Northwest e-Health’. Internationally, in order to bring global benefit from sharing e-epidemiological data, methods and expertise, open standards are needed. We recommend that public health informatics groups come together to define global standards, or ‘meta-standards’, for inter-operable e-Epidemiology.

We propose a federated model for e-Labs, using open standards-based technologies that are easy to deploy and maintain. Each local node of the federation will own systems with integrated datasets. Running a research protocol across multiple nodes will use a technology we call ‘research objects’. This approach promotes local ownership, enhancement and use of data; it matches existing NHS information governance arrangements and importantly avoids the need to build a centralised data warehouse.

The federated model permits a unique pseudonymisation scheme to be used for each data linkage request, thus reducing the feasibility of performing a brute force attack on identifiers such as NHS number.

The interface between the NHS network and the public Internet is unidirectional. We propose that the NHS e-Lab engine will call out to the e-Lab research portal to pull in new dataset requests. Once prepared, new datasets will be pushed back to the research portal, and the user is notified.

The process of creating a new linked-dataset begins with the user planning the data required using the metadata catalogue. A new Research Object is populated with the metadata of the required data and sent to the NHS e-Lab linking engine. The Link Controller then coordinates the linking of data from all the target sources (both clinical and non-clinical), gathering a pseudonymised data set from each. When the pseudonymised data are returned they are linked together, then anonymised before being released to the research portal. In exceptional circumstances, where secondary linkage of transformed data is required, the researcher works with pseudonymised data, within the healthcare organisation, under its information governance.

Each healthcare organisation that participates in the e-Lab is able to modify the data they expose and the security policy that controls data-release at any time. This ensures local ownership of the data by the data controller. The metadata for the dataset are published to the central e-Lab where they are used for planning data queries.

Within the e-Lab federation there are three tiers of data sensitivity. The healthcare providers maintain identifiable data and only release pseudonymised data to the NHS e-Lab. The NHS e-Lab is able to link at the individual record level using pseudonymous identifiers. Once the linking has been completed, the data are anonymised, before being delivered to the researcher using the research portal. This ensures that no identifiable data are released by the data controller and that no pseudonymised data leaves the safety of the NHS.

The e-Lab research portal gives the scientist access to data, workflows, computing resources and research objects. An aggregated metadata repository describes the data available across all the e-Lab nodes. This enables users to plan their data set request and determine the data that will be returned in accordance with their security privileges.