From online banking to biobanking: designing and implementing a data delivery platform for researchers of the China Kadoorie Biobank

Sansome, S[1]

[1] University of Oxford

Introduction

As a large prospective cohort study of 512,891 participants, we routinely integrate data via linkage to population health outcomes to deliver up-to-date research datasets. Building on experience from the private sector, we have implemented a platform that supports the delivery and audit of secure datasets to internal and external researchers.

Objectives and Approach

We aimed to create a platform that delivers secure research datasets for preliminary analyses and fieldwork with dynamic censor dates. It must also provide multiple static versions of an analysis-ready database with fixed censor dates. Individual participant outcome data from population health sources (death and disease registries and health insurance agencies) can be integrated and linked regularly to other health-related data, e.g., genetic, bioinformatics, and medical device. With knowledge of data management strategies used in major financial institutions, we have produced systems that successfully implement the techniques of data warehousing, multiple concurrent environments and secure dataset access and delivery.

Results

As at the end of 2017, we had over 300 registered and approved researchers eligible to request datasets. Using our platform, we have recorded over 150 requests and successfully delivered over 100 de-personalised and encrypted datasets to external researchers around the world. In addition, we have supplied secure datasets to over 20 Global Health MSc and DPhil students studying at The University. The platform currently hosts 4 completed analysis-ready databases with censor dates ranging from 31st December 2013 to 11 years of follow-up as of 31st December 2016. A 5th analysis-ready database (with the most recent outcome data from participants' death, disease and hospitalisations) is already under development. We plan during 2018 to make available more data sources and outcome data to our external researchers.

Conclusion/Implications

We have developed a versatile platform that delivers secure datasets for researchers from a selection of analysis-ready databases, each with differing censor dates and available data sources. This platform is scalable and can accommodate regular integration of known follow-up data sources along with new and emerging data sources.