Exploring the joint analysis of routine data and pathogen genomic datasets in the investigation of outbreaks of gastrointestinal infection

Area of research interest: Foodborne pathogens
Project code: FS301019
Conducted by: University of Warwick

Background

During a foodborne disease outbreak, accurate and rapid identification of the source of the outbreak is critical so that measures can be taken as quickly as possible to prevent further cases of disease. Recent technological developments have made it much cheaper and faster to obtain genomic sequence data, with the result that sequencing of pathogen genomes can potentially be used to great effect during outbreak investigations: the use of pathogen genomic sequence data in surveillance can complement other datasets generated during investigation of foodborne disease incidents, and has the potential to improve the detection and investigation of outbreaks.

However, there are challenges – for example, data sharing across public and private sector organisations can be difficult to achieve, particularly during an acute incident. This project used expert review of outbreak case studies to identify:

1. benefits achievable by the effective integration of genomics and other datasets in incident and outbreak investigation;
2. issues and barriers to this integration;
3. learning points from the reviewed outbreaks;
4. examples that could support the motivation of wider partners to work together to establish data sharing priorities.

Research Approach

This was a scoping project comprising technical pilot work and an expert workshop to provide a framework for developing methods, infrastructure, and partnerships to bring available data and technologies to the investigation and of control foodborne outbreaks.

The approach was to:

1. Use literature review, a review of a national outbreak database, and contact with experts to select a pilot collection of 15 previously investigated outbreaks.

2. Seek full reports and data from any accompanying epidemiological studies conducted within Public Health England (PHE) and full details of any laboratory characterisation of the available isolates
3. Identify and seek access to datasets relevant to the investigations held by external partners, testing data access and governance issues with a range of partners.
4. Prepare a case study report on each outbreak, including identification of opportunities for integrating genomics optimally, opportunities for big data approaches, and issues identified such as technical and information governance requirements.
5. Review these case reports in an expert workshop including PHE, FSA, and academic group expert staff to (i) develop a protocol for prospective capture of this type of multi-dimensional data in outbreaks, (ii) identify priorities area for development, (iii) identify issues that need to be addressed to support collaborative work across organisations and datasets.

Results

Genomic cluster detection

1. Genomic clusters appear to detect foodborne outbreaks with good specificity and can allow earlier detection if sequencing is rapid and appropriate small clusters are investigated. Investigation of even small clusters may efficiently identify sources for small or early outbreaks and is recommended where resources allow.
2. Cluster investigation is substantially more efficient where complementary forms of data are available. Obtaining at least basic case epidemiology data on a consistent and accessible basis nationally is the first priority for complementary information.
3. Widely distributed (spatially and temporally) outbreaks are identified more commonly with the addition of pathogen genome sequencing to surveillance. Complementary case epidemiology data to support this outbreak type (e.g. shared brands or suppliers) should be considered in decisions on what data to collect for each case.

Epidemiological control data for comparison of e.g. exposures with cases

4. Routine sources of population exposure data may provide a valid alternative or complement to control data from outbreak studies and should be explored for use in outbreaks. This requires mapping of available data and establishing access so that it is available rapidly in outbreak settings.
5. Internet panel controls are increasingly used and very efficient. They are inevitably a biased population sample and protocols for use should consider mitigations.
6. Gaining timely access to good population exposure representative controls could be facilitated by pre-arrangement of access to more representative controls. This would require agreements and a protocol analogous to the developments already made for internet panel controls.

Increased descriptive and analytical epidemiology accuracy and efficiency and sequencing data

7. Integrating sequencing data into case definitions supports more efficient and informative application of trawling questionnaires and food tracing studies.
8. Where appropriate, increasing the specificity of case definition using sequencing data can allow more efficient and accurate case-control studies.

Linking cases to upstream sources in the food chain

9. Linking human cases to possible sources using sequence data may be very useful. However, especially in the absence of corroborative epidemiological information, linkage to a reservoir or source by sequence data alone could be incorrect if the outbreak clone is relatively stable and widespread.

Both structured samples of such sources (animal, food and environment) and the assembly from other potentially available data, such as data from testing done by industry for their internal quality control, are high priorities to support more robust inference from sequence data in the investigation of foodborne disease.

Research report

England, Northern Ireland and Wales

PDF
[View Exploring the joint analysis of routine data and pathogen genomic datasets in the investigation of outbreaks of GI infection as PDF(Open in a new window)](#) (1.62 MB)