Supplementary material - Classifying literature mentions of biological pathogens as experimentally studied using natural language processing

Appendix A. Availability of data and materials

The code has been made publicly available from a GitHub repository.

- Code developed in the project to build the pathogen characterisation data set is available from: `https://github.com/READ-BioMed/readbiomed-ncbi-pathogen-dataset-generation`
- Code for the pathogen characterisation methods is available at: `https://github.com/READ-BioMed/readbiomed-pathogen-annotator`.
- The MTIMLExtension package used to train text classifiers with SVM and AdaBoostM1 are available from: `https://github.com/READ-BioMed/MTIMLExtension`.

In addition to code, we have made available from GitHub data used in this project, which includes the data sets that we have created using NCBI resources.

- The data sets created for this project have been made available from: `https://github.com/READ-BioMed/readbiomed-pathogens-dataset`.

Appendix B. Regular expressions for toxin identification

The regular expressions below have been tested with the Java programming language for the identification of toxins in MEDLINE citations.

(?i)(Abrus.*abrin.*toxin)
(?i)(Aflatoxin)
(?i)(Anatoxin-A)
(?i)(Batrachotoxin)
(?i)(Botulinum.*toxin)

(?i)(Brevetoxin)
(?i)(Ciguatoxin)
(?i)(Conotoxin)
(?i)(decarbamoylsaxitoxin)
(?i)(Fusariotoxins.*(T-2))
(?i)(gonyautoxin)
(?i)(Maitotoxin)
(?i)(Mycotoxin)
(?i)(neosaxitoxin)
(?i)(Palytoxin)
(?i)(Ricinus.*ricin.*toxin)
(?i)(saxitoxin)
(?i)(Staphylococcus.*enterotoxin—Enterotoxin)
(?i)(Tetrodotoxin)

Appendix C. Examples of actively research pathogens vs not actively researched

We show information about two MEDLINE citations to highlight the difference between pathogen mentions that discuss viruses actively researched in the study discussed in a citation and pathogens mentioned in a citation that is not the actively studied pathogen.

The first one (PMID: 19040284) discusses work about the *hepatitis delta virus*. Samples of the virus were collected and analysed.

The second one (PMID: 27756212) mentions the *Japanese encephalitis virus (JEV)* in the context of studying its vaccine for the dengue virus. In this paper, JEV is referenced in the citation but it is not the pathogen of study in the paper.

PMID: 19040284

Title: Molecular epidemiology of **hepatitis delta virus (HDV)** in Iran: a preliminary report

Abstract: To identify hepatitis delta virus (HDV) genetic variability and its circulating genotypes amongst infected Iranian patients, 25 patients with positive anti-HDV status from different parts of Iran were enrolled in this cross-sectional study. A portion of the HDV delta antigen was amplified, sequenced, and subjected to molecular and phylogenetic analysis. Clinical features and virological markers were evaluated. HDV RNA could be detected in 88% of anti-HDV positive cases (22 patients) with chronic hepatitis B virus (HBV) infection and liver cirrhosis. Phylogenetic analysis revealed that all Iranian patients were infected by genotype I (clade 1) of HDV, supported by a high bootstrap value (100%, 1,000 replicates). All HDV-positive patients were coinfected with genotype D1 of HBV. No significant association was determined between demographic, clinical, and virological variables in the population studied. In conclusion, the present molecular epidemiology survey reveals that clade 1 of HDV is predominant among coinfected HBV patients in Iran.

PMID: 27756212

Title: Japanese encephalitis vaccine-facilitated dengue virus infection-enhancement antibody in adults

Abstract:

Background: Dengue virus (DENV) and Japanese encephalitis virus (JEV) belong to the genus Flavivirus, and infection with a virus within this genus induces antibodies that are cross-reactive to other flaviviruses. Particularly in DENV infection, antibodies to DENV possess two competing activities: neutralizing activity and infection-enhancing activity. These antibody activities are considered central in modulating clinical outcomes of DENV infection. Here, we determined the neutralizing and infection-enhancing activity of DENV cross-reactive antibodies in adults before and after JE vaccination.

Methods: Participants were 77 Japanese adults who had received a single dose of inactivated Vero cell-derived JE vaccine. A total of 154 serum samples were obtained either before or approximately a month after a single dose of JE vaccination. The antibody-dependent enhancement (ADE) activity to each of four DENV serotypes and the neutralizing activities to DENV and to JEV were determined in each of the serum samples by using baby hamster kidney (BHK) cells and FcγR-expressing BHK cells.

Results: A total of 18 post-JE immunization samples demonstrated cross-reactivity to DENV in an anti-DENV IgG ELISA. DENV neutralizing antibodies were not detected after JE vaccination in this study. However, undiluted post-JE vaccination serum samples from 26 participants demonstrated monotypic and heterotypic ADE activity to DENV. ADE activity was also observed in 1:10-diluted samples from 35 of the JE vaccine recipients (35/77, 45%).

Conclusion: In summary, JE vaccination induced DENV cross-reactive antibodies, and at sub-neutralizing levels, these DENV cross-reactive antibodies possess DENV infection-enhancement activity. The results also indicate that cross-reactivity to DENV is associated with high levels of JEV neutralizing antibodies and, the DENV cross-reactivity is further facilitated by JE vaccination.