Multimedia Appendix 1. Data Characterisation form using during full text review.

Data Characterisation	Additional Information
Topic	State topic of paper: ILI = Influenza Related Illness FBI = Food Borne Illness IID = Infectious Intestinal Disease Disease H1N1 Public Health
Country	State country of data collection
Primary data type	State data type: Twitter Online restaurant review Other
Control data type	State if gold standard measure was used for comparison, e.g.: PHE foodborne disease outbreak data Centre for Disease Control ILI Network
Keyword selection	State how keywords were generated for data collection: Single Keyword Keyword List Manually generated Automatically generated Knowledge based generation (mining blogs / webpages etc.)
Methods	State methods used for identifying cases of disease / ailment Basic keyword matching Keyword / lexicon based analysis (term frequencies etc.) Semantic analysis (based on non-textual features) Machine learning classifier (SVM, NB etc.) Clustering / topic modelling (LDA etc.) Custom classifier State performance of methods: Precision and recall. For supervised classifiers: Size of training dataset Process of generation (Manual labelling / Amazon Mechanical Turk etc.)
Results	State results of incidence calculation: Correlation coefficients (if calculated)
Demographic analysis	State if the paper discussed the demographic limitations of using CGD for disease / ailment surveillance.
Limitations	Limitations of the study as outlined by the author, e.g. sample size. Limitations of the study as identified by reviewer.