ISDS 2019 Conference Abstracts

Use of N-grams and Term Relationship Graphs in the Syndrome Definition Development Process

Nimi Idaikkadar, Nelson Adekoya, Aaron Kite-Powell, Achintya N Dey

Centers for Disease Control

Objective

To describe the use of uni-grams, bi-grams, and tri-grams relationships in the development of syndromic categories.

Introduction

The use of syndromic surveillance systems has evolved over the last decade, and increasingly includes both infectious and non-infectious topic areas. Public health agencies at the national, state, and local levels often need to rapidly develop new syndromic categories, or improve upon existing categories, to enhance their public health surveillance efforts. Documenting this development process can help support increased understanding and user acceptance of syndromic surveillance. This presentation will highlight the visualization process being used by CDC's National Syndromic Surveillance Program (NSSP) program to develop and refine definitions for syndromes of interest to public health programs.

Methods

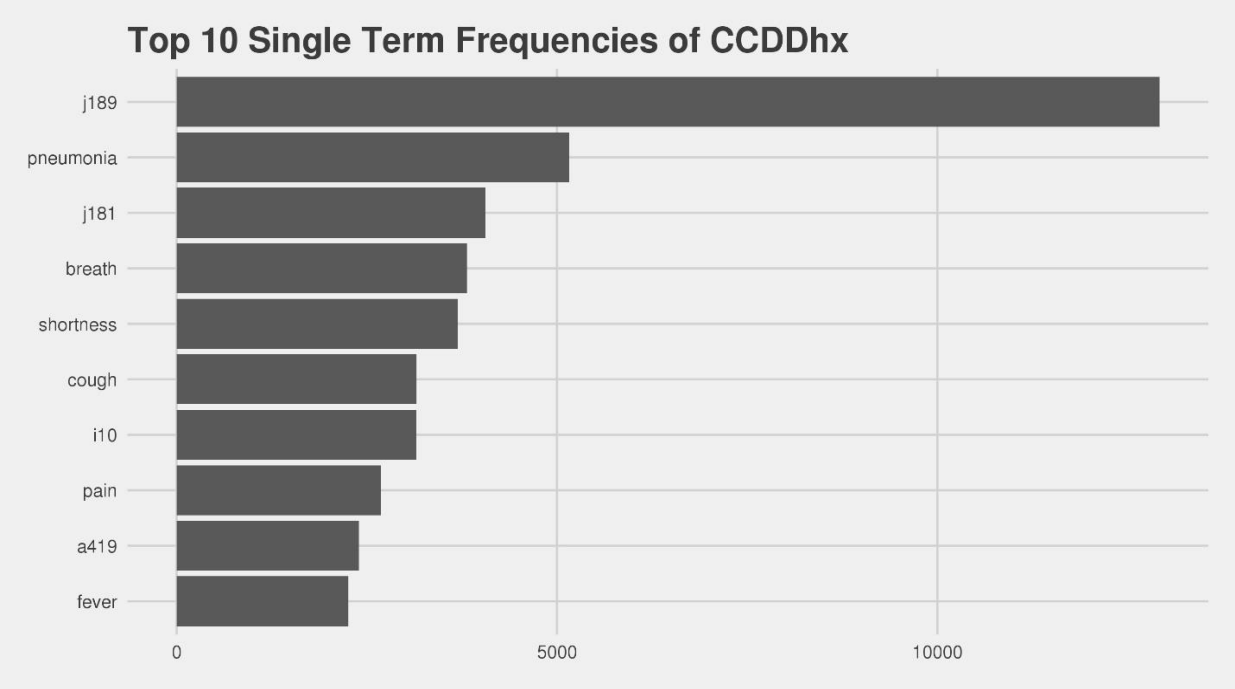
Development of a syndromic definition is an iterative process that starts with an analyst testing how different terms, which are assumed to be associated with the topic of interest, and diagnostic codes are noted in the chief complaint and discharge diagnosis code fields. The analyst then manually scans through the resulting line list of patient chief complaint text and diagnostic codes to determine whether the query terms match the intended syndromic concept. Typically, more terms and diagnostic codes are then added to the query using Boolean operators, and other terms are negated and removed. To facilitate summarization of the resulting terms and diagnostic codes CDC's NSSP program developed programs with R that extracted data using the ESSENCE application programming interface (API), and the chief complaint query validation data source (CCQV). We use N-gram analysis, which is extensively used in text mining, to show co-occurrences of words in a consecutive order. The co-occurrences of words can be a uni-gram which represents a single word, bi-gram for two words, and tri-grams for three words. The process tokenizes the chief complaint text and diagnosis code fields, with some pre-processing of the text and removal of stopwords. Uni-grams, bi-grams, and tri-grams are then calculated for the top 200 combinations along with term and diagnostic code co-occurrence. Other visualizations that can be used are network graphs, which show the connections between different chief complaints terms and also between discharge diagnosis codes and chief complaint terms. The use of these graphs provides an insight into the frequency and relationship between terms and codes.

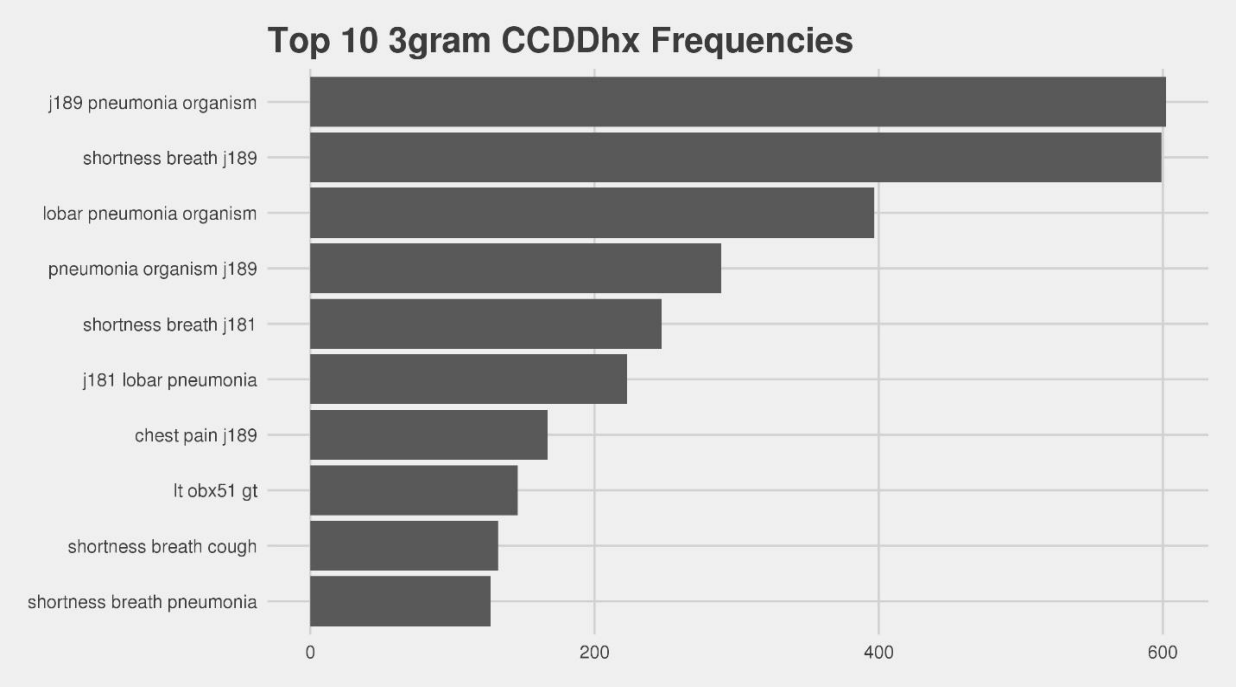
Results

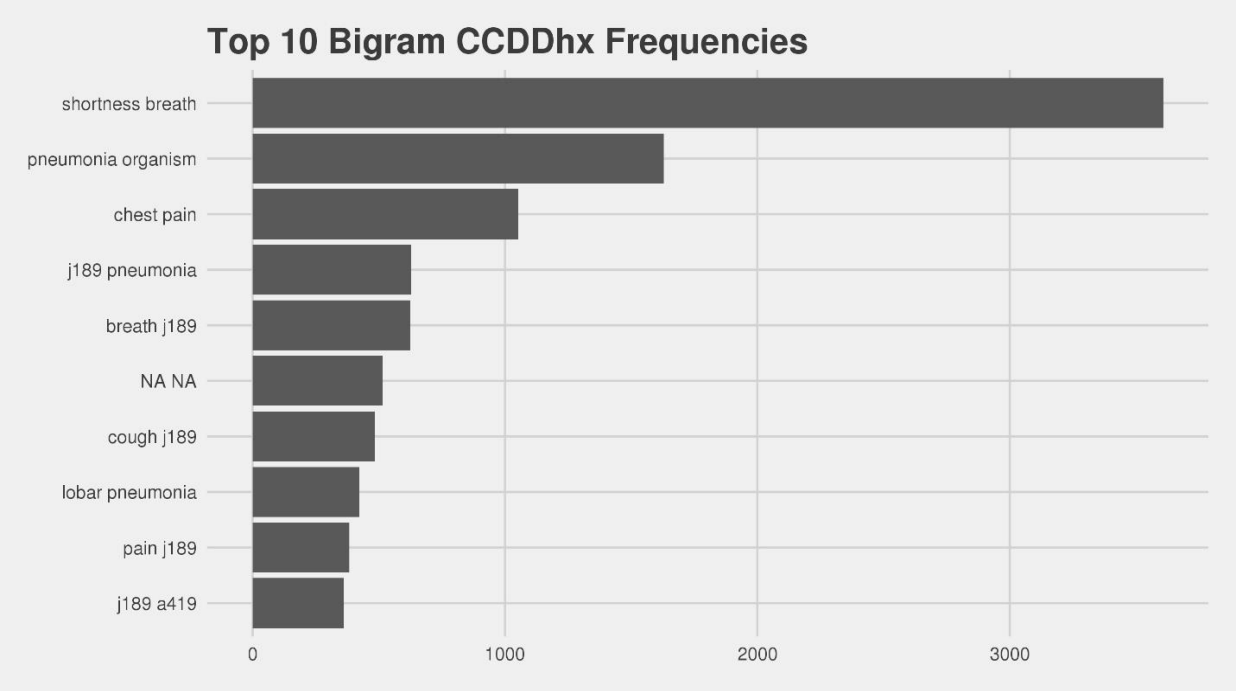
To support the development of new syndrome definitions we used the R program to produce two time series graphs. The first time series graph is used to show the volume of visits over the user's indicated time period and the second shows the median chief complaint compared over the user's indicated time period. A series of histograms showing frequency of the uni-gram, bi-grams, and tri-grams are also used during the development process. Lastly, two network diagrams are used to show the co-occurrence between term and diagnostic codes. The use of this range of graphs during the syndrome definition development process provides multiple ways to view the characteristics of the chief complaint and discharge diagnosis fields. The sample graphs below can be used by the analyst to illustrate key information.

Conclusions

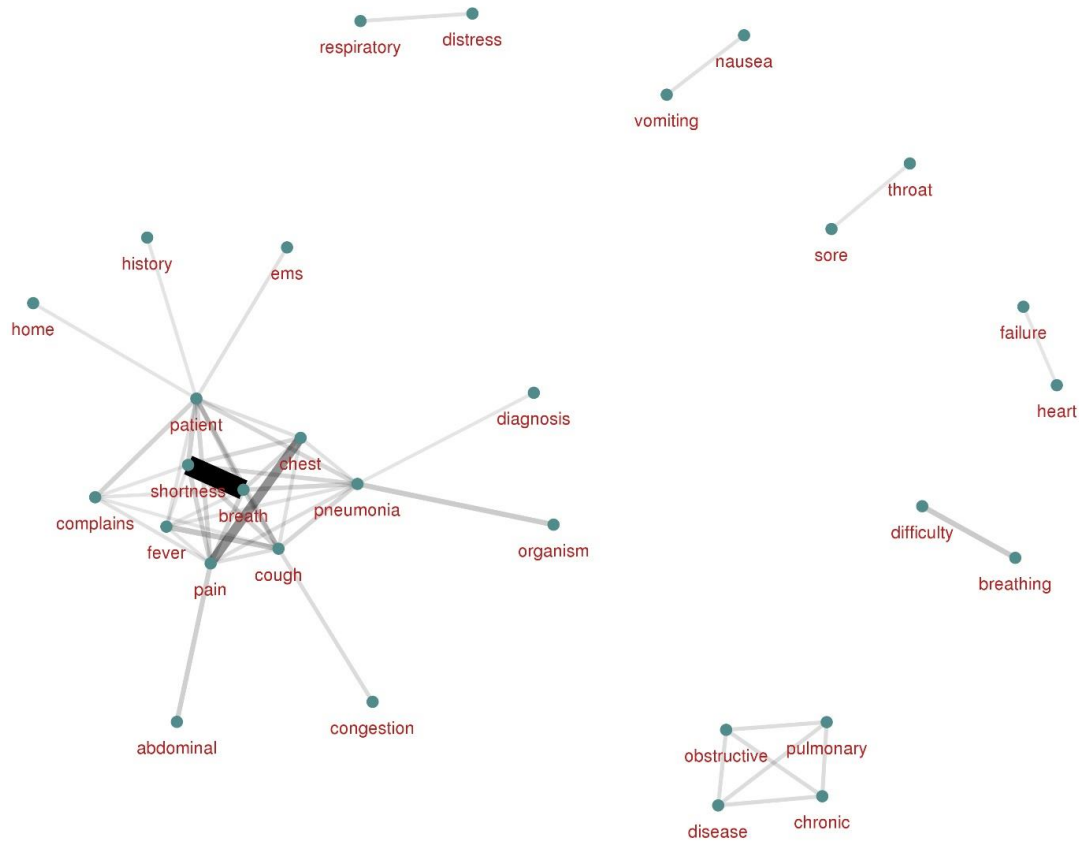
Through this development process and the use of graphs the relationship between the syndrome definition and search terms can be visualized. In addition when using this process, the analyst could be specific as to the syndrome of interest or be broad, allowing a generic trend series monitoring of the syndrome. The search words can also be based on specific local or regional terms and the relationship terms set to include or exclude certain terms. Use of this process for the development of syndrome definitions can support the use of syndromic surveillance and offer the opportunity to further refine the process. After the syndrome has been developed, the analyst can consider spatial or temporal analysis.







Chief Complaint Term Relationship Graph - Top 50



Chief Complaint and Discharge Diagnosis History Term Relationship Graph - Top 50