

# DADAR: A New Situational Awareness Platform for Data Analysis, Detection and Response

Dave Carter\* and Joel D. Martin

National Research Council Canada, Ottawa, ON, Canada

## Objective

To develop an adaptable platform for periodically loading semi-structured medical text, extracting syndromic information using advanced natural language processing, detecting outbreaks in the data (including the ability to tune sensitivity vs. specificity on a syndrome-by-syndrome basis so as to reduce the rate of false alarms), generating timely cartographic surveillance reports, and providing tools to quickly validate or rule out syndromic alerts.

## Introduction

We previously experimented with tracking influenza in ER chief complaint data using existing syndromic surveillance tools. We identified several deficiencies in these tools: poor natural language processing, inefficient user interfaces, frequent (thus costly) false alarms, and one-size-fits-all approaches to syndromes.

Furthermore, we were surprised that some epidemiologists we spoke with had relatively little faith in existing surveillance tools, and so we set out to build one that would address their concerns: DADAR (Data Analysis, Detection, And Response).

## Methods

We designed and implemented a system to perform four key tasks:

1. ingest data periodically (nightly); clean it as required (i.e., perform medical spell checking); and extract symptoms, infer syndromes from symptoms, and classify syndromes by example using text mining techniques
2. identify aberrations (outbreaks) in the data using any of the common existing algorithms (CDC EARS, RLS, SaTScan, moving average) in a more trustworthy fashion
3. generate and email single-page PDF alert summaries (with map) for each aberration found
4. offer a web-based investigation tool so that the recipient of an alert (an epidemiologist) can quickly and easily confirm or rule out the alert, and, if necessary, increase or decrease the confidence threshold (user-selectable false alarm rate) of that alert going forward

Our system is data agnostic. Data models are automatically built for a given data set the first time it is imported. All geotemporal data that can be flattened (represented in a single spreadsheet) is easy to load into DADAR. Geocoding and GIS visualization tools are built-in. We map both case counts and counts per 10 000 households.

Our system can detect different kinds of aberrations, not just rapid increases in counts over a short period of time. For example, we can send alerts when an unusually low number of cases is detected for a given definition (vaccination uptake, for example).

We offer a web-based rapid hypothesis investigation dashboard to:

- verify or deem invalid the alerts that our system sends out
- drill down into alerts' symptoms for preliminary differential diagnosis
  - form and validate "what if" syndrome definition scenarios using complex boolean logic
  - set up nightly/weekly/monthly informational reports and/or periodic alerts based on such ad-hoc syndrome definitions
  - quickly answer questions about emerging outbreaks on-the-fly, on-the-go

This dashboard updates a map, a flexible graph, and a list of relevant cases as the user changes date ranges and/or boolean search criteria.

We took care to make sure that the math behind our alerts is trustworthy. In cases where we generate multiple hypotheses (i.e., combining pairs, triples, quadruples, and so on, of geographically adjacent zip codes), we correct for multiple hypotheses. Our advanced natural language processing improves our ability to correctly count cases in the first place.

We built our platform so that we could rapidly integrate our latest research. New aberration detection algorithms and new natural language processing components can be plugged in easily. Software-savvy users can develop and plug in their own intellectual property using web services.

## Results

We have launched a release candidate of our software with a provider of telehealth services in two Canadian provinces, ingesting detailed data for as many as 4500 nurse triage calls per day. Each call record contains demographic data, call metadata, and detailed nurses' notes in free text. We extract symptoms and syndromes automatically. We currently monitor calls for pockets of influenza-like, gastrointestinal, and respiratory illness; and we regularly perform ad-hoc reporting for exigent concerns like measles or pertussis outbreaks.

We have been able to generate relevant reports or assuage concerns about pertussis, ILI, and E. coli outbreaks much faster than our partner's existing tools.

## Conclusions

We found room to improve on existing syndromic surveillance tools. We built and deployed a new and extensible surveillance platform to address common criticisms of older tools and increase users' confidence in syndromic alerts.

## Keywords

surveillance; situational awareness; data fusion; outbreak detection; informatics

\*Dave Carter

E-mail: david.carter@cnrc-nrc.gc.ca

