**GRITS**

Global Rapid Identification Tool Set for Disease Threat Identification and Diagnostics

Zachary S. Gold, Andrew G. Huff, Amy M. Slagle, Nathaniel Breit, Russell S. Horton

March 2015

Partners: ISID, Kitware, ProMed-mail, epidemic, HealthMap

Funder: DTRA

***Executive Summary***

GRITS is a biosurveillance application that allows users to more efficiently monitor textual data sources for infectious disease threats by conducting an automated initial analysis that succinctly summarizes the critical epidemiological information in a text source, and proposes possible diseases associated with a report. GRITS can be accessed through a web-interface for detailed analysis of single text samples. Through the web-interface users can access dynamic visualization of GRITS analyses, perform powerful queries of an index of over 250,000 disease reports for similar disease occurrences, and examine historical disease emergence events pertinent to a text sample. The comprehensive GRITS API can be used to continuously analyze information feeds, and large collections of data. The API also enables GRITS technology to be easily incorporated into larger surveillance systems. GRITS is built on robust and flexible Natural Language Processing (NLP) and machine learning algorithms that can be modified for specific purposes. In conjunction with human expertise GRITS will be a valuable tool for disease surveillance.

***Background***

Infectious diseases pose a significant threat to global health,[[1]](#footnote-1) and economic stability.[[2]](#footnote-2) Due to extensive globalization, infectious diseases can spread at unprecedented rates. Failure to detect localized infectious disease threats early can have pandemic consequence, as demonstrated by Severe Acute Respiratory Syndrome (SARS) in 2003,[[3]](#footnote-3) and H1N1 influenza in 2009.[[4]](#footnote-4) Unfortunately, in many resource-poor regions, most prone to infectious disease threats,[[5]](#footnote-5) disease surveillance remains sparse, and is improving slowly, if at all.[[6]](#footnote-6) This problem was illustrated recently in the ongoing Ebola Virus epidemic in West Africa, in which the etiological agent of disease was not identified until 85 days after the index case.[[7]](#footnote-7) More effective global disease surveillance is desperately needed.

Traditional disease surveillance relies on the transmission of epidemiologic information through public health infrastructure. This robust system leverages enormous epidemiologic expertise, extensive public health communication networks, and increasingly sophisticated laboratory testing, to detect, report and identify disease outbreaks. Despite the power of traditional disease surveillance, it is often hampered by the heterogeneous quality of regional healthcare infrastructure, international discrepancies in reporting procedures, and the influence of political and economic interests.[[8]](#footnote-8)

The burgeoning field of Digital Disease Detection (DDD) uses computer and Internet technologies to complement traditional disease surveillance. DDD tools are developed to help solve a diverse array of surveillance problems. Commonly, DDD tools apply NLP and machine learning algorithms to perform automated disease threat surveillance of traditional and non-traditional data sources. Prominent examples of DDD tools include the Global Public Health Intelligence Network (GPHEN)[[9]](#footnote-9), PulseNet,[[10]](#footnote-10) HealthMap,[[11]](#footnote-11) Argus,[[12]](#footnote-12) BioCaster,[[13]](#footnote-13) and the Global Early Warning System for Major Animal Diseases Including Zoonoses (GLEWS).[[14]](#footnote-14)

***Problem***

DDD is an innovative, but innately challenging field. Although, the quantity of digital information is immense, the signal to noise ratio is very low. To curate data sources for epidemiologically significant information, like disease or spatial content, many DDD tools require substantial human resources. This dependency may be rate limiting and makes it difficult to provide disease detection that is superior to traditional surveillance systems. Conversely, some DDD tools make extensive use of NLP or machine learning algorithms to predict disease risk, but do not incorporate sufficient epidemiological, or clinical expertise. This has led to inaccurate predictions of disease risk with potentially dangerous implications (9). DDD tools that do not circumvent health analysts, and enable public health analysts to examine digital sources more efficiently are critically needed.

***Solution***

GRITS is a biosurveillance tool created to monitor and analyze textual data sources (e.g., news articles, medical reports). GRITS extracts critical epidemiologic information (e.g., case-counts, symptoms, pathogens, transmission types, hosts, dates, locations) from text and delivers ranked differential diagnoses, helping make rapid infectious disease threat detection possible. GRITS also visualizes disease occurrence on a timeline and map, and provides links to similar and potentially related disease reports using a customizable search feature. GRITS is available through a web-application for detailed analysis of specific text, and through an API for rapid analysis of large collections of text, and for integration with other applications. GRITS is a flexible, pluripotent tool that can be tailored to address other surveillance needs.

***Intended Audience***

GRITS is intended to serve as a tool for public health and military analysts in their daily disease surveillance efforts. GRITS will serve both individual users, through the web interface, and developers of other surveillance applications, through a comprehensive diagnostic API.

***Components and Features***

**GRITS Web-Application**

The GRITS web-application is a dynamic web-interface for performing detailed analysis of a text sample. Through the web-application, user can submit a text sample to GRITS for analysis and exploration, and can view the previous GRITS analyses they have conducted. Users can view a dynamic visualization of GRITS analyses through the Diagnostic Dashboard, use the Find Similar Articles Feature to conduct customizable searches of a pre-analyzed article index for related reports, and search for relevant disease emergence events using the Find Similar Disease Emergence Events feature.

*Diagnostic Dashboard*

The GRITS Diagnostic Dashboard provides users with a ranked differential diagnosis of potential diseases associated with submitted text. Diagnoses are determined using a MaxEnt BoW classifier trained on disease labels HealthMap assigned to articles over a recent 2-year interval. Possible diagnoses are ranked based on confidence score. The keywords that contribute to each diagnosis and their relative weights can be inspected from within the diagnostic dashboard (Fig. 1).

The GRITS Diagnostic Dashboard identifies, and succinctly visualizes the important epidemiological information in a text sample. This includes disease related keyword (disease, pathogens, symptoms, modes of transmission, and hosts), dates, locations and case counts. Extracted information is highlighted in an annotated article view that can be filtered by category (Fig. 2). Locations are plotted in a map view, and a histogram of temporal information is shown in a timeline view (Fig 3). Date extraction is done via the Standford SUTime library. The Geoname resolution uses a custom algorithm with geonames.org data. Case count extraction uses the CLiPS Pattern library’s search module with a number of search phrases tailored to meet GRITS requirements.

*Find Similar Articles Feature* (Fig. 4)

The GRITS Find Similar Articles Feature provides a simple, customizable interface to query over 250,000 disease related articles collected by HealthMap over a 2-3 year interval. These articles have all been subjected to GRITS analysis. Articles can be searched by any combination of the following information obtained from their completed GRITS analysis; diagnosed disease, extracted keywords, publication date range, and/or country. Additionally, a query of articles similar to the text sample is automatically conducted based on the results of the text sample’s GRITS analysis. Search results can be viewed spatially using Kitware’s geojs library, or in a list view. The list view includes links to the original articles, and selected meta-data, including diagnosed disease, case counts, distinct keywords and more. Aggregations of publication dates and countries associated with search results are visualized in histograms.

*Find Similar Disease Emergence Events Feature*

This feature allows users to query the Emerging Infectious Disease Repository (EIDR) developed by EcoHealth Alliance for disease emergence events that may be related to a text sample. Development of EIDR is ongoing. As of March 2015, EIDR contained information on 369 infectious disease emergence events occurring between 1940 and 2013.[[15]](#footnote-15) GRITS automatically generates a list of emergence events associated with the diagnosed disease(s) for a text sample. EIDR events can also be queried by host, pathogen, disease, transmission mode, country and date. Aggregations of emergence event start dates and associated countries are visualized in histograms. Users can examine emergence events in detail through the EIDR web-application, which can be accessed by clicking on any emergence event search result.

**GRITS API**

The GRITS API allows users to apply GRITS diagnostics continuously on large collections of data and to use GRITS intelligence in any application. API users can submit documents for analysis through a REST API and receive comprehensive results in a JSON format, including the differential disease diagnosis, identified pathogens, hosts, symptoms, locations, dates and more. Extracted keywords are annotated with their occurrences in the original text, so they may be easily highlighted in a user interface, indexed for search, used as features for statistical evaluation or machine learning, or subjected to further analysis. This open and flexible API allows developers to easily integrate GRITS analysis capabilities into their own application with simple calls to a central web service, with no requirement to set up or maintain their own installation of GRITS, although that is supported as well.

***Value***

GRITS allows users to more efficiently monitor textual data sources, including non-traditional sources, for infectious disease threats. In the hands of the astute surveillance expert, GRITS could be used as a powerful tool for disease surveillance. The NLP and machine learning technology that GRITS is built on is robust and flexible. GRITS could be expanded to incorporate a customizable triaging system that curates text sources temporally, spatially, by diagnosed disease, or by epidemiological keyword. Additional ontologies could be created to train GRITS to make educated conclusions on additional complex variables besides disease, like pathogen class, report risk level, or the existence a novel pathogen. An alert system could be built into GRITS to warn users of potentially dangerous clusters of reports. Additionally, through the GRITS API, the tool could be incorporated into larger surveillance systems, like the Defense Threat Reduction Agency’s prototype Biosurveillance Ecosystem, and run continuously on those systems data feeds.[[16]](#footnote-16) Through the web-interface users can evaluate the functionality, and epidemiologic foundation of GRITS. This increases the transparency of the tool, and invites scrutiny, which should lead to its refinement, and continued development.

GRITS is fundamentally sophisticated NLP and machine learning software that has been tailored to address disease surveillance needs. The GRITS technology could be applied to a plethora of topics. For instance, GRITS could be of value to the financial sector. Textual sources are rich with indicators of investor sentiment, and are often monitored by NLP tools.[[17]](#footnote-17) GRITS could be tailored to detect clusters of investor sentiment indicative of an emerging financial crisis, or market shift.

1. Morens DM, Folkers GK, Fauci AS (2004) The Challenge of Emerging and Re-emerging Infectious Diseases. Nature. Vol. 430:242-249 [↑](#footnote-ref-1)
2. Fonkwo PN (2008) Pricing Infectious Disease. EMBO Reports. Vol. 9:S13-S17 [↑](#footnote-ref-2)
3. Fleischaurer AT (2003) Outbreak of Severe Acute Respiratory Syndrome – Worldwide, 2003. CDC-MMMR Vol. 52:226-228 [↑](#footnote-ref-3)
4. (2010) The 2009 H1N1 Pandemic: Summary Highlights, April 2009-April 2010. http://www.cdc.gov/h1n1flu/cdcresponse.htm [↑](#footnote-ref-4)
5. Jones KE, Patel NG, Levy MA, et al. (2008) Global Trends in Emerging Infectious Diseases. *Nature*. Vol. 451:990-993 [↑](#footnote-ref-5)
6. Chan EH, Brewer TF, Madoff LC, et al. (2010) Global Capacity for Emerging Infectious Disease Detection. Proceedings of the National Academy of Sciences. Vol. 107:21701-21706 [↑](#footnote-ref-6)
7. World Health Organization (2015) Global Alert and Response. One Year into the Ebola Epidemic: A Deadly, Tenacious and Unforgiving Virus. http://www.who.int/csr/disease/ebola/one-year-report/introduction/en/ [↑](#footnote-ref-7)
8. Sturevant JL, Aranka A, Brownstein JS (2007) The New International Health Regulations: Considerations for Global Public Health Surveillance. Disaster Medicine and Public Health Preparedness. Vol. 1:117-121 [↑](#footnote-ref-8)
9. Mykhalovskiy E, Weir L (2006) The Global Public Health Intelligence Network and Early Warning Outbreak Detection: A Canadian Contribution to Global Public Health. Canadian Journal of Public Health. Vol. 97:42-44 [↑](#footnote-ref-9)
10. http://www.cdc.gov/pulsenet/ [↑](#footnote-ref-10)
11. Brownstein JS, Freifeld CC, Reis BY, Mandl KD (2008) Surveillance Sans Frontières: Internet-based emerging infectious disease intelligence and the HealthMap project. PLoS medicine. Vol. 5:e151 [↑](#footnote-ref-11)
12. Hartley DM, Nelson NP, Walters R, et al. (2010) Landscape of International Event-Based Biosurveillance. Emerging Health Threats Journal. Vol. 3:e3 [↑](#footnote-ref-12)
13. Collier N, Doan S, Kawazoe A, et al. (2008) BioCaster: Detecting Public Health Rumors with a Web-based Text Mining System. Bioinformatics. Vol. 24:2940-2941 [↑](#footnote-ref-13)
14. http://www.glews.net/ [↑](#footnote-ref-14)
15. Gold ZS, Huff AG, Tilchin C et al. (2015) EIDR Emerging Infectious Disease Repository: A Repository and Web Application Dedicated to the Emergence of Infectious Disease. [↑](#footnote-ref-15)
16. Cleary ME, Antelman ET, Markuzon N, et al. (2014) COMBS: A Biosurveillance Ecosystem (BSVE) Prototype. Online Journal of Public Health Informatics. Vol. 6 [↑](#footnote-ref-16)
17. Lugmayr A, Friedrichsen M (2013) Predicting the Future of Investor Sentiment with Social Media in Stock Exchange Investments: A Basic Framework for the DAX Performance Index. Media In *Handbook of social media management value chain and business models in changing media markets*. Heidelberg: Springer.

    [↑](#footnote-ref-17)