

Title: Evaluation and Verification of the Global Rapid Identification of Threats System (GRITS) for Infectious Diseases in Textual Data Sources

Corresponding author:

Andrew G. Huff, Ph.D.
Michigan State University
736 Wilson Road
East Lansing, MI 48824
Ph#: 612-743-1265
Email: andrewgeorgehuff@gmail.com

Co-Authors in order:

Nathan Breit, B.S.
EcoHealth Alliance
New York, NY, USA

Toph Allen, M.P.H.
EcoHealth Alliance
New York, NY, USA

Karissa Whiting, B.A.
EcoHealth Alliance
New York, NY, USA

Christopher Kiley, Ph.D.
Defense Threat Reduction Agency
Fort Belvoir, VA, USA

Keywords: natural language processing; biosurveillance; emerging infectious disease; intelligence; NLP; EID surveillance; biosurveillance methods;

Acknowledgement

This study was made possible by the generous support of Defense Threat Reduction Agency (DTRA) through a contract (Contract No. HDTRA1-13-C-0029) awarded to EcoHealth Alliance. The contents are the responsibility of the authors and do not necessarily reflect the views of DTRA or the United States Government.

ABSTRACT

Global Rapid Identification Tool Set (GRITS) is a biosurveillance application that enables infectious disease analysts to monitor non-traditional information sources (e.g., social media, online news outlets, ProMED-mail reports, and blogs) for infectious disease threats. GRITS analyzes these textual data sources by identifying, extracting, and succinctly visualizing epidemiologic information and suggests potentially associated infectious diseases. This manuscript evaluates and verifies the diagnoses that GRITS performs and discusses novel aspects of the software package. Via GRITS' web-interface, infectious disease analysts can examine dynamic visualizations of GRITS' analyses and explore historical infectious disease emergence events. The GRITS API can be used to continuously analyze information feeds, and the API enables GRITS technology to be easily incorporated into other biosurveillance systems. GRITS is a flexible tool that can be modified to conduct sophisticated medical report triaging, expanded to include customized alert systems, and tailored to address other biosurveillance needs.

INTRODUCTION

Infectious diseases pose a significant threat to global health and economic stability.[1,2] Due to extensive globalization and urbanization, infectious diseases can spread at unprecedented rates.[3] Small and localized infectious disease threats can rapidly become international catastrophes, as demonstrated by influenza (H1N1A) in 2009, Ebola Virus Disease in 2014, and Middle Eastern Respiratory Syndrome in South Korea and the Middle East.[4,5,6] Identifying infectious disease outbreaks is critical to reducing overall harm and preventing epidemics. Increasing biosurveillance systems' detection and communication speed may contribute to a reduction in overall health and economic consequences from infectious diseases.

Traditional biosurveillance systems rely predominantly on local clinicians, laboratory technicians, and public health practitioners to identify infectious disease outbreaks. In part, these systems identify cases via routine patient care where samples are collected. Then, laboratory testing is performed on the collected samples, and clinical case definitions are established. Via these processes, infectious disease cases are typically reported to a centralized authority, which then aggregates and monitors cases for signs of an above normal caseload.

Unfortunately, traditional biosurveillance systems are limited by their cost, their limited geographic coverage, and their inability to rapidly communicate results. For example, an upgrade to the existing United States' Biowatch program cost \$61 million dollars and was canceled before its completion. Furthermore, effective biosurveillance systems depend on the quality of underlying health care infrastructure, which can be highly variable geographically. Without a vast network of healthcare infrastructure,

traditional biosurveillance systems may not be sensitive enough to detect rare infectious diseases. Also, some traditional biosurveillance systems may not be accurate (e.g., lack of laboratory capacity); thereby, overwhelming the infectious disease analyst with incorrect or irrelevant information.

In part, these barriers lead to incomplete geographic coverage, varying by disease type, for traditional surveillance systems. As a direct result of incomplete geographic coverage for some infectious diseases, infectious disease outbreaks in regions with inadequate healthcare infrastructure may not be identified in the outbreak's early stages, as seen with the ongoing Ebola Virus Disease epidemic in West Africa (not identified as EVD until 85 days after the first case). Governments, who may be reticent to announce an outbreak for fear of economic harm, often control traditional biosurveillance systems as occurred with the 2003 SARS outbreak. With current healthcare technology and investment levels traditional biosurveillance systems lack complete coverage.

Typically, traditional biosurveillance systems are tailored to a single infectious disease (e.g., ILInet, Malaria Early Warning System, European Legionnaire's Disease Surveillance Network), requiring clinicians to report diseases based on pre-defined lists. [7] Different governing entities have different lists of infectious diseases that must be reported by clinicians, and these lists are at times updated to reflect the current needs of the public health community. In some cases, traditional biosurveillance capabilities are implemented for specific classes of diseases, transmission pathways, and specialized laboratory capabilities (e.g., ILInet, Foodnet, Pulsenet). Most traditional biosurveillance systems are well suited to monitor known infectious disease threats

(e.g., poliovirus, influenza) but are not designed to detect threats from unknown or extremely rare infectious diseases.[8]

The term syndromic surveillance is used to refer to a number of different types of biosurveillance systems where symptoms are used to classify the type of infectious diseases.[8] Syndromic surveillance was first used to describe biosurveillance cases that conformed to particular clinical case definitions (this is especially useful when monitoring diseases where no laboratory test exists). However, its usage has expanded to encompass most forms of biosurveillance outside of traditional biosurveillance systems. These include systems that collect information on hospital admissions, pharmaceutical sales, employee absenteeism, and other data streams that are used to detect outbreaks.[8]

Digital disease detection, also called digital biosurveillance, refers to analysis of web data for insight on public health and infectious disease systems.[8] The term is broadly defined to include various uses of web-native information: (1) aggregation of medical reports from subject matter experts (e.g., ProMED-mail); (2) computational models built upon search results and web traffic (e.g., Google Flu Trends), and, (3) models built on other clusters of search terms around infectious disease trends. Digital biosurveillance examines indirect evidence for infectious disease cases (e.g., textual data sources from symptomatic people) and must work in combination with traditional biosurveillance methods. Digital biosurveillance's greatest potential is that it can possibly identify potential outbreaks where traditional biosurveillance systems do not exist and can rapidly detect and communicate infectious disease outbreaks.

Digital biosurveillance holds promise, but has yet to fulfill a concrete role as an early warning system in public health biosurveillance. There is disagreement about the utility of digital disease surveillance for predicting influenza outbreaks.[9,10] Initially, there was some evidence that Google Flu Trends was useful in forecasting developing influenza outbreaks;[8] however, Google Flu Trends was inconsistently accurate from year to year and there were substantial flaws in Google Flu Trends ability to predict regular seasonal influenza peaks and irregular pandemic influenza.[10] Furthermore, tools and analytical methods that rely upon human curation of data feeds (e.g., ProMED-mail, HealthMap) require significant human capital and appear to scale with the amount of training and education of the human curators.[11]

Despite these weaknesses of digital disease surveillance, Natural Language Processing (NLP) is a potentially useful tool for biosurveillance systems. NLP is able to give structure to unstructured textual data. For example, NLP has been used to automatically classify electronic medical records (EMR) from emergency rooms into categories for syndromic biosurveillance,[12,13] especially in cases where specific clinical definitions are scant (e.g., invasive mold).[14] In the realm of digital biosurveillance, efforts are underway to apply NLP to social media streams.[15] Using NLP to systematically create structured data from unstructured text may enable the monitoring of innumerable local sources of infectious disease information globally. Digital biosurveillance methods that use NLP may lead to the accurate and rapid detection of infectious disease outbreaks in places where traditional biosurveillance systems are insufficient. For these reasons, EcoHealth Alliance developed the Global

Rapid Identification of Threats (GRITS) that uses NLP to identify emerging infectious disease threats in textual sources.

METHOD

GRITS uses natural language processing to determine which infectious diseases are most likely associated with an input text sample. Articles are processed using a combination of NLP methods to identify disease-related features from the text. These features are passed to an ensemble of binary logistic regression classifiers, which work together to “diagnose” the article, ranking diseases by predicted probability.

GRITS’ search function

GRITS presently searches an index of over 250,000 infectious diseases related articles. Elasticsearch assigns relevance scores to individual terms using TF-IDF (term frequency-inverse document frequency) based models, which weight matches according to how common words are in a document divided by how rare they are across the corpus of documents. Additionally, GRITS’ extracted feature metadata for each article, including infectious disease keywords, date, and location, are searchable can be used to sort search results.

Feature extraction

GRITS’s text processing and NLP algorithms, written mainly in Python, extract disease-related and contextual features from texts, and store these features as annotations on the text. The algorithms are detailed in Supporting Information, and code samples are provided (Supplemental 1). Prior to analysis, non-English text is translated using Bing Translator.

Feature extraction is performed using Python's standard pattern-matching libraries and the NLTK package to match keywords from a variety of compiled ontologies of terms related to infectious disease and public health (Table 1).

Ontology	Contents	Description
Biocaster Ontology	General disease ontology	English terms for symptoms, diseases and pathogens are used as features
GRITS Ontology	Curated ontology of symptoms, control measures, descriptions of infected individuals, diseases, disease categories, environmental factors, hosts, host uses, modes of disease transmission, occupations, disease risks, vectors, and zoonotic types	Collection of keywords and terms gathered and vetted from a consensus of experts at EcoHealth Alliance
HealthMap Disease Labels	Diseases identified as significant by HealthMap and used for their disease labels	Used as outcome in logistic regression models
The Disease Ontology	Human disease related terms, phenotypic characteristics, medical vocabulary disease concepts	Disease names and synonyms are used as keyword features. Predicates from disease definitions
USGS Topographic Feature Vocabularies	Environmental factors	Subset used as features (all labels and synonyms of type owl#Thing)
Wordnet	English language ontology that maps word relatedness	Hyponyms and lemmata for a set of epidemiology-related root keywords are used as features

Table 1. The ontologies used in GRITS, their contents, and their descriptions.

Features are categorized: Diseases, Pathogens, Symptoms, Hosts, and Modes of Transmission. Dates are extracted with the Stanford SUTime Java library. Locations are matched with a custom algorithm that uses data from the GeoNames database in addition to a number of heuristics to reduce false positive matches. Case Counts are identified using the CLIPS Pattern library's search module, with a number of specifically

tailored search phrases. GRITS stores extracted features in JSON with information about their position in the text, so they can be viewed in separate from the document or in their original context.

Classifier training, verification, & evaluation

GRITS uses the binary relevance method (as implemented in scikit-learn's `sklearn.multiclass.OneVsRestClassifier`) to predict the disease referred to by a body of text. This uses an ensemble of logistic regression classifiers, one for each disease label (approximately 120). Each classifier estimates the probability that a text passage is associated with a single disease, given the vector of features extracted by GRITS's NLP algorithms. Multiple occurrences of features were not counted.

Classifier training and testing used a randomly selected corpus of approximately 150,000 articles from a 2–3 year period (of the 250,000 article set), and collected and assigned a single disease label each by analysts. Classifiers were trained on a subset of approximately 12,000 articles. Each classifier fit a logistic regression model, using articles with that classifier's disease label as positive responses and all other articles in the training set as negative responses.

RESULTS

GRITS' diagnostic performance evaluation

The classifiers' performance was tested over a set of approximately 3500 health news articles and ProMED reports. A confusion matrix was composed, from which the micro-averaged F1 score was calculated across all classifiers. The micro-averaged F1 score sums all true positives, false negatives and false positives, evaluating classifier

performance across all diseases in the GRITS ontology. To determine the relative contribution of features for a given diagnosis on a text, the regression coefficients for each classifier are rescaled to sum to 1, then multiplied by the estimated probability of that disease for that text.

GRITS' diagnostic algorithm verification

The GRITS disease classification system has an overall precision (positive predictive value) of 64% and recall (sensitivity) of 63%. The overall F1 score is 0.317. However, GRITS diagnoses some diseases very well (**Table 2**) and some diseases very poorly (**Table 3**). These results included translations and were not skewed due to translation.

Disease	Precision (PPV)	Recall (Sensitivity)	F1 score	N of positive articles
Avian Influenza	0.923	0.932	0.928	208
Hepatitis	0.989	0.866	0.923	112
Influenza	0.905	0.959	0.931	830
Listeriosis	0.921	0.951	0.936	62
Measles	0.931	0.964	0.947	226
Polio	0.893	0.976	0.933	43
Salmonella	0.871	0.983	0.924	124
Scabies	1	0.862	0.925	29
Syphilis	0.928	0.928	0.928	14
Tuberculosis	0.939	0.951	0.945	82

Table 2. GRITS' top 10 performing disease classifications with > 10 positive in the testing set. “

Disease	Precision (PPV)	Recall (Sensitivity)	F1 score	N of positive articles
Rubella	1	0.09	0.166	11
Respiratory Illness	0.666	0.181	0.285	11
Campylobacter	0.875	0.538	0.666	13
Chikungunya	0.651	0.823	0.727	34
<i>Clostridium difficile</i>	0.888	0.235	0.372	34
Eastern Equine Encephalitis	0.833	0.416	0.555	12
Hemorrhagic Fever	0.486	0.947	0.642	19
HIV/AIDS	0.687	0.733	0.709	15
Lyme Disease	0.588	0.833	0.689	12
<i>Neisseria meningitidis</i>	0.707	0.659	0.682	44

Table 3. GRITS' bottom 10 performing disease classifications with > 10 positive in the testing set.

DISCUSSION

Context for biosurveillance

GRITS provides a framework for classifying the infectious disease-related content in potentially arbitrary text. Monitoring digital disease signals for impending infectious disease threats means that biosurveillance capacity can be extended to areas where the healthcare and public health infrastructure is inadequate. This is crucial since many emerging infectious disease threats occur in places where traditional biosurveillance infrastructure is scant.

In the hands of the astute public health analyst, GRITS is a powerful tool for infectious disease biosurveillance that allows users to efficiently monitor non-traditional data sources for infectious disease threats. It can extend the capabilities of analysts to triage and monitor a wider range of textual sources, increasing coverage of non-traditional digital disease surveillance in areas where traditional systems do not exist, and supplementing traditional methods where they do. GRITS is currently used in the Defense Threat Reduction Agency's (DTRA) Biosurveillance Ecosystem (BSVE) to identify infectious disease threats globally.

Limitations and future directions

Large sources of annotated disease-related textual data, required to accurately train machine learning classifiers, is uncommon, difficult to come by, and time-consuming to create. The HealthMap data used to train the GRITS classifiers is sufficiently large, but each article is only labeled with one disease, even when a text may mention multiple diseases. This means that disease traits extracted from an article may not map specifically to the disease that article is labeled with, negatively impacting classifier training.

The HealthMap training data consists largely of aggregated online news articles, WHO, and ProMED-mail reports. These texts have a set of features linguistic properties specific to online news articles related to health. If GRITS were applied to other sources of text, like scholarly articles or social media feeds, new sets of training data would likely have to be curated.

In an active surveillance system using GRITS, feature ontologies and article classifiers should be updated on an ongoing basis. New diseases will emerge, disease

classifications and ontologies will change, and as the way diseases described in public discourse will change, and the GRITS system must be updated to prevent diminishing accuracy. Incorporating feedback from GRITS users (from the results of individual articles) into classifier training would improve classifier fit for that article type.

GRITS currently exists as a standalone web application. However, its utility would be increased as part of a larger suite of biosurveillance tools, and with connections to continuous data feeds. These would enable building out various decision support capabilities around the GRITS toolset. For instance, GRITS could store processed text sources and display summaries of articles temporally, spatially, or by diagnosed disease or public health keyword. A alert system could be built on top of this dataset to warn users of potentially dangerous clusters of reports, and additional ontologies could be created to train GRITS to make educated conclusions on additional complex variables like pathogen class, report risk level, or the emergence of a novel pathogen. Additionally, through the GRITS API, these tools are planned for incorporation to the Defense Threat Reduction Agency's Biosurveillance Ecosystem (BSVE), and will run continuously on BSVE data feeds.[16]

REFERENCES

- 1 Morens DM, Folkers GK, Fauci AS. The challenge of emerging and re-emerging infectious diseases. *Nature* 2004;**430**(6996):242-249
- 2 Fonkwo, PN. Pricing infectious disease. *EMBO Reports* 2008;**9**(Suppl 1):S13-S17.
- 3 Hosseini P, Sokolow, SH, Vandegrift KJ, et al. Predictive power of air travel and socio-economic data for early pandemic spread. *PLoS ONE* 2010;**5**(9):e12763.
- 4 Fraser C, Donnelly CA, Cauchemez S, et al. Pandemic potential of a strain of influenza A (H1N1): early findings. *Science* 2009; **324**(5934):1557-1561.
- 5 Baize S, Pannetier D, Oestereich L, et al. Emergence of Zaire Ebola virus disease in Guinea. *N Engl J Med* 2014;**371**(15):1418-1425.
- 6 Schar D, Daszak P. Ebola economics: the case for an upstream approach to disease emergence. *EcoHealth* 2014;**11**(4):451-452.
- 7 Pearson B, Sy F, Holton K, et al. Fear and stigma: the epidemic within the SARS outbreak. *Emerg Infect Dis* 2004;**10**(2):358-363.
- 8 Morse SS. Public health surveillance and infectious disease detection. *Biosecur Bioterror* 2012;**10**(1):6-16.
- 9 Dugas AF, Jalalpour M, Gel Y, et al. Influenza forecasting with Google flu trends. *PLoS ONE* 2013;**8**(2):e56176.
- 10 Olson DR, Konty KJ, Paladini M, et al. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. *PLoS Comput Biol* 2013;**9**(10):e1003256.
- 11 Wei CH, Harris BR, Li D, et al. Accelerating literature curation with text-mining tools: a case study of using PubTator to curate genes in PubMed abstracts. *Database* 2012; bas041.
- 12 Friedlin J, Grannis S, Overhage JM. Using natural language processing to improve accuracy of automated notifiable disease reporting. *AMIA Annu Symp Proc*; 2008; **2008**:207-211.

- 13 Doan S, Conway M, Phuong TM, et al. Natural language processing in biomedicine: a unified system architecture overview. In: Trent, R, ed. *Clinical Bioinformatics*. New York: Springer 2014:275-294.
- 14 Ananda-Rajah MR, Martinez D, Slavin MA, et al. Facilitating surveillance of pulmonary invasive mold diseases in patients with haematological malignancies by screening computed tomography reports using natural language processing. *PLoS ONE* 2014;**9**(9):e107797.
- 15 Dredze M, Paul MJ. Natural language processing for health and social media. *IEEE Intelligent Systems* 2014;**29**(2):64-67.
- 16 Biosurveillance Ecosystem factsheet.
http://www.dtra.mil/Portals/61/Documents/CB/BSVE%20Fact%20Sheet_04282015_PA%20Clear.pdf Accessed Aug 2015.