Title: Passive methods to map environmental factors to non-communicable disease

Authors: Kristin Isaacs,[1] Chris Grulke,[1] Raina Brooks,[2] Madeline Reich,[3] Ryan Edwards,[4] Michael-Rock Goldsmith[1*]

Keywords: infoveillance, webidemiology, big data, public health, natural language processing, search-term volume analysis, named-entity recognition

Affiliations:
1. US - Environmental Protection Agency, Office of Research & Development, RTP NC, 27711
2. Student Services Contractor at U.S. EPA, RTP NC
3. Fuquay-Varina High School, Fuquay-Varina, NC 27526
4. North Carolina State University, 2200 Hillsborough St, Raleigh, NC 27695

***Corresponding author**: Michael-Rock Goldsmith

Mail Code: E205-01, 109 T.W. Alexander Drive, Research Triangle Park, NC 27711
Phone: 919-541-0497
Email: goldsmith.rocky@epa.gov

1. Introduction

Many pervasive and multifactorial disorders of modern non-communicable and non-infectious diseases (i.e. obesity, asthma, migraine, autism) need to be better understood and explored. The cost of these disorders to the healthcare system and the general public's quality of life are two great reasons to study a variety of chronic or acute health conditions that have been orphaned or underfunded. Recently, medical and public health researchers have begun to take advantage of the huge increase in the amount of information available via the internet and the concomitant advance in data science technology to explore potential disease-related factors and the temporal and spatial relationships among them at a variety of population levels.[1]

An individual's disease/health status can be thought of as a complex balance of a variety of biological, environmental and lifestyle factors (Figure 1). Quantifying these factors for large numbers of individuals can elucidate meaningful trends and aid in hypothesis generation. Historically, our knowledge of lifestyle and environmental factors and their interactions have been limited. We are now using new methods to collect disease factor information from publically-available web-based or social media sources. These methods are based on syndromic surveillance via search term analysis followed by extraction of disease-relevant information (including mentions of symptoms, treatments, activities, events, consumer product use, and other related human behaviors) from social media data streams. The information gleaned using these new methods can be explored on both spatial and temporal scales and compared with other available public data (ex. census, public health, housing, weather, environmental data) to try and better understand, quantify, and visualize the factors involved in

disease etiology. In this article, we (1) describe the methods (2) provide some recent simple case studies of this type of approach from our group and others (3) formalize the approach using a "Look, Listen, and Learn" paradigm, and (4) provide a useful listing of analysis tools and data sources.

2. Methods and Examples

Search Volume Analysis

Search volume analysis involves a quantitative investigation of the aggregate search term volume from a variety of search engines such as Google or Yahoo, using publically available tools such as Google Trends. These approaches provide temporal and spatial information about what people are searching for, and hypothetically, what they are doing, eating, or buying. This approach can be fine-tuned to validate the assumption that there is a correlation between human activity/behavior and information-seeking patterns by comparing global, national, or local epidemiological statistics (prevalence/incidence) and provides a rational starting point to identify the optimal seasons and locations to administer surveys with maximum value of information. The search-volume approach does not assist in directly identifying the demographic information of a population although it can be combined with census data to “glean” or infer population demographics of different geographic locations.

The key step in search volume analysis is the development of appropriate search terms for a given research question (for example, a set of disease symptoms). Text analysis (i.e. wordclouds), set-generation methods (such as Google Sets), and crowdsourcing approaches such as Mechanical Turk (https://www.mturk.com) can be useful in developing meaningful sets of search terms.

Google Trends analyses have recently been used to examine the temporal properties of Lyme disease outbreaks.[2] In addition, a simple example of a disease-related search volume analysis is shown in Figure 2. We performed a Google Trends analysis of the term "migraine" in 40 different languages. The global, geographically-specific relative search-volume results were compared against migraine prevalence for 14 different global regions as reported by the World Health Organization (WHO). Despite the simplicity of the search term, a significant correlation could be seen. This example could be refined to target specific countries or symptoms, or to correlate these results to other global data (such as weather).

We have also recently used search volume analysis in an attempt to elucidate trends in consumer product use for the purpose of risk assessment of chemicals. We have determined in preliminary studies that there is a high degree of correlation between the Google Trend search volumes associated with a set of terms describing different types of cleaning products or personal care products and the numbers of those products actually found in homes in field studies of consumer product use. This finding will allow us to extrapolate our methods to predict product use for product categories for which we have little or no data.

Mining of Social Media (Microblogging) Streams

Another approach to exploring symptoms and related factors is to deal directly with micro-blogging "fire-hose" such as Twitter or Tumblr. These microblogging environments can provide similar information to the search-volume approach but require the development of natural language processing (NLP) methods, such as named-entity recognition (NER) term tokenizer that identifies the key taxonomy for a given domain of interest. Appropriate taxonomies are required to more correctly or accurately identify and capture representation of terms related to a specific symptom, disease, human activity, or behavior.[3] However, an advantage of these methods is that they can provide immediate correlations between symptoms and activities via tracking search terms in the same individuals. Geographic information can be collected as well, as microblog entries (e.g. Tweets) may be geocoded based on user settings. Finally, it is easier to extract demographic information for individuals from these data via user-provided profiles or from information within the microblog streams themselves.

One useful method we are developing is the mapping of microblog entries to the generalized activity and location codes described in the U.S. Environmental Protection Agency's Consolidated Human Activity Database (CHAD). This database is used in EPA's assessments of exposure to chemicals. It has already been shown that NLP can be used to translate spoken word diaries directly into CHAD codes,[4] and we are extending this to text-based diaries. This will allow us to map microblog entries directly to activities and locations that are relevant to pollutant exposures, and then correlate these with other disease symptoms or other relevant factors. To this end we have also begun taxonomy development related to CHAD activities in a simple 9-person “how would you tweet that” experiment.

There have been numerous attempts to mix both search-volume derived data and micro-blogging data into the mainstream of predicting, extrapolating and anticipating human behavior as it relates to disease and epidemic outbreaks, such as Google Flu Trends[5] (http://www.google.org/flutrends/us/#US) or Health Map (http://healthmap.org/en/). These attempts have found this method to be quite useful, as is why using this method in an effort to alleviate the big data problem of elucidating disease etiology of diseases, both non-communicable and non-infectious, seems plausible. Our proposed approach for combining these methods for the purpose of informing hypothesis development is described below.

3. The Look, Listen, and Learn Paradigm

We now attempt to combine and formalize and formalize the above methods. We have found a "Look, Listen, and Learn" paradigm (illustrated in Figure 3) extremely useful in framing our approach to new questions. We can start from a global perspective, using aggregated data that contains temporal and spatial information related to specific keyword search terms and work our way down to population, community and eventually individual level, where we can collect detailed information from social media data streams. This paradigm can be summarized with the following steps:

(1) Look: Problem identification and key factor term development

The step starts by defining target public health problem (e.g. asthma, migraine, obesity, psoriasis). A scoping analysis of the literature (pubmed.org) is performed to identify keywords for search formulation and inventory information that could potentially to be used later for method validation and ground-truthing. In addition, "open-health" websites (such as curetogether.com) can be explored for potential symptom/treatment related search terms. Text analyses or visualizations (such as wordclouds) can be useful in extracting important terms from large volumes of text (for example abstracts).

(2) Listen: Aggregate search term volume analysis at different spatial/geographic and temporal scales

This step involves using search term volume analysis (for example, via Google Trends) at global (multi-lingual), population (monolingual), and community levels to elucidate potential temporal and geographical trends and identify target locations/populations of interest.

(3) Learn: Target populations and times of interest for detailed social media analysis leading to hypothesis generation

This final step includes fine-tuning the analysis of any key identified locations and seasons via NLP mining of anonymized geo-coded social streams (microblogs). In this way, more detailed information can be explored (e.g. hourly as opposed to seasonal patterns, demographic details gleaned from social media profile, other activities or factors gleaned from the data stream).

Many of the factors explored in any of the above three steps (e.g. disease symptoms or relevant activities) can be mapped against other available data to elucidate data gaps, visualize temporal/seasonal variation in disease prevalence, and explore correlations among potential factors. A selection of freely available spatial/temporal data and some useful tools are provided in Tables 1 and 2.

This paradigm can be used for scoping analysis and hypothesis generation. Advantages and disadvantages of this approach are discussed in Table 3. While these methods will never take the place or more traditional methods in epidemiology, they have the potential to inform the design of future focused studies that optimize selection of geographic and seasonal sampling patterns, relevant cohorts, and studied factors.

Disclaimer:

The United States Environmental Protection Agency through its Office of Research and Development funded and managed the research described here. It has been subjected to Agency review and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Acknowledgements: Drs. Peter Preuss, Stacy Katz, Gail Robarge, Kevin Kuhn and other members of the Innovations team of the US-Environmental Protection Agency's Office of Research and Development provided support for this research. The Shaw University Research Internship Program made it possible to obtain the contributions from M. Reich and R. Edwards.

References

1. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. J Med Internet Res. 2009 Mar 27;11(1):e11.

2. Seifter A, Schwarzwalder A, Geis K, Aucott J. The utility of "Google Trends" for epidemiological research: Lyme disease as an example. Geospat Health. 2010 May;4(2):135-7.

3. Paul, M.J. , Dredze, M. “Your are What you tweet: Analyzing Twitter for Public Health” http://www.cs.jhu.edu/~mdredze/publications/twitter_health_icwsm_11.pdf

4. Guinn C., and Reeves, D. J., Using a Spoken Diary and Heart Rate Monitor in Modeling Human Exposure to Airborne Pollutants for EPA’S Consolidated Human Activity Database. In: G.A. Uzochukwu et al. (eds), Proceedings of the 2007 National Conference on Environmental Science and Technology, Springer Science, NY, 2009.

5. Carneiro HA, Mylonakis E. Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clin Infect Dis. 2009 Nov 15;49(10):1557-64.

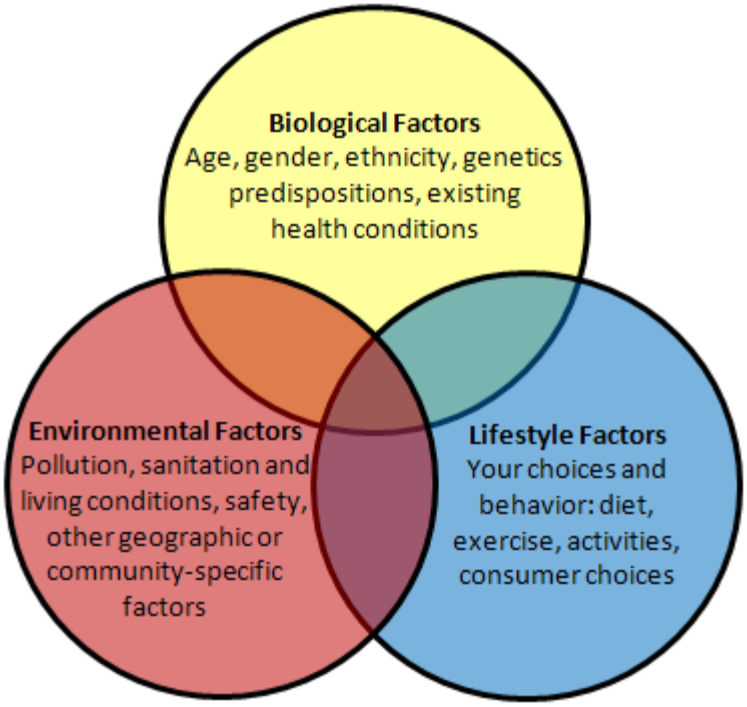
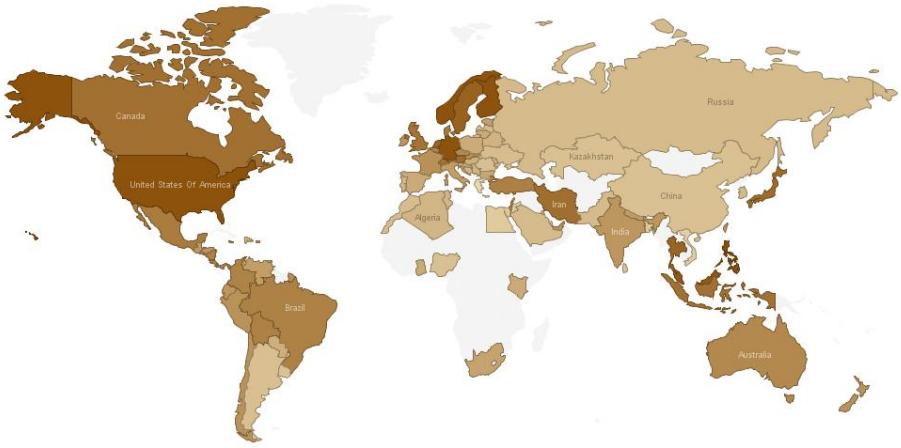


Figure 1: Overlapping domains that give rise to an individual's health status, being a complex function of biological, environmental and lifestyle factors. Being able to capture electronically information of this type voluntarily "journaled" or "streamed" by individuals has tremendous value in public health modeling and methods



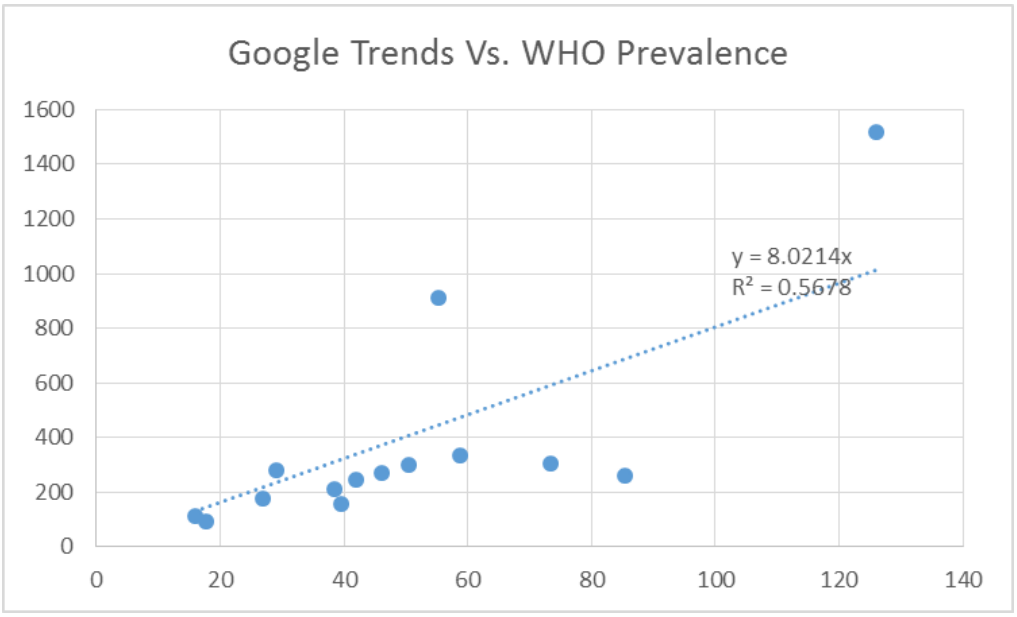


Figure 2: Comparison of global relative search volumes for "migraine" compared to WHO migraine prevalence data.

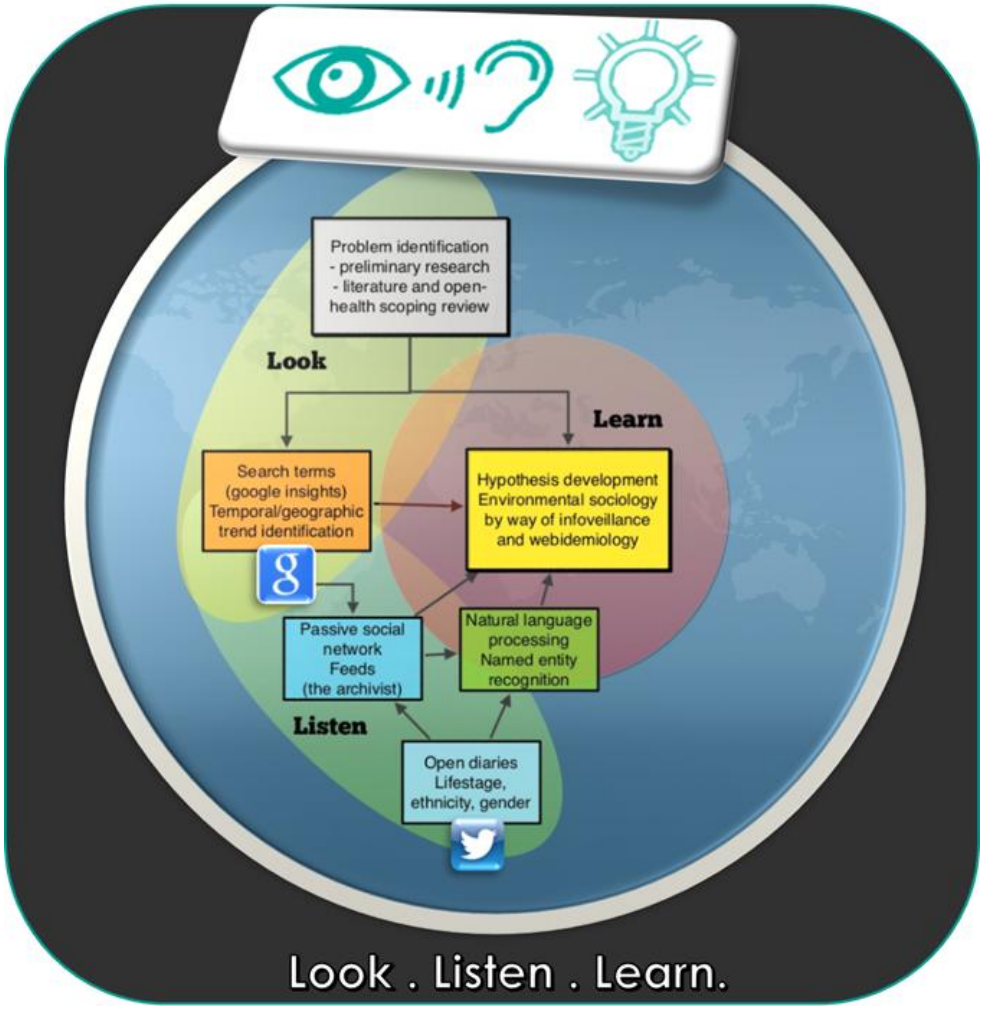


Figure 3: The look, listen, and learn paradigm.

Table 1: Some useful data streams by type, source, and accompanying open-access resource URL. Most of these contain geographically-specific information.

Data type	Source	URL
Asthma Statistics	The global burden of asthma: executive summary of the GINA Dissemination Committee Report	http://onlinelibrary.wiley.com/doi/10.1111/j.1398-9995.2004.00526.x/full
Pollen Data	American Academy of Allergy, Asthma & Immunology	http://www.aaaai.org/global/nab-pollen-counts.aspx
Air Quality	Air Now	www.airnow.gov
U.S Census	U.S. Census Bureau	http://www.census.gov/main/www/access.html
National Air Toxics Assessments	U.S. EPA	http://www.epa.gov/nata/
American Housing Survey (Housing Information)	U.S. Census Bureau	http://www.census.gov/housing/ahs/
Diet and SES (Food Environment Atlas)	Economic Research Service	http://www.ers.usda.gov/data-products/food-environment-atlas/go-to-the-atlas.aspx#.UZ4keKLqnxA
Weather Forecast	National Weather Service - NOAA	http://www.weather.gov/forecastmaps
List of WHO Regions and Sub regions	World Health Organization (WHO)	http://www.who.int/entity/healthinfo/statistics/gbdestimatesregionallist.xls
Migraine Statistics	World Health Organization (WHO)	http://www.who.int/healthinfo/statistics/bod_migraine.pdf
PM 2.5 map of Earth (served as comparison)	National Aeronautics and Space Administration (NASA)	http://www.nasa.gov/topics/earth/features/health-sapping.html

Weather/ Activity Relationship (served as comparison)	National Aeronautics and Space Administration (NASA)	http://icp.giss.nasa.gov/education/urbanmaap/projects/projects_asthma5.html
Human Activities (Consolidated Human Activity Database)	US EPA	http://www.epa.gov/heasd/chad.html
Human Activities (American Time Use Survey)	Department of Labor	http://www.bls.gov/tus/

Table 2: Useful tools for performing passive web-based analysis of disease factors

Tool	Definition of tool	How tool was used	URL	Reference for definition
Cure Together	Free public site that allows people around the world to share quantitative information about their medical conditions, including symptoms and treatments that worked best for them.	Used to identify self described symptoms and treatments.	http://curetogether.com/	http://curetogether.com/blog/about/
Google Maps	A map service that you view in your web browser and that provides geocodes.	Used to retrieve geocodes.	https://maps.google.com/	http://support.google.com/maps/bin/answer.py?hl=en&topic=1687350&safe=on&answer=144352
Google Scholar	Freely accessible web search engine that provides a	Used to search for literature, in order to get an idea of factor	http://scholar.google.com	http://www.google.com/intl/en/scholar/about.html

	simple way to broadly search for scholarly literature across many disciplines and sources.	and information landscape of disease in question.		
Google Sets	A tool within Google Drive that identifies keywords that are semantically related.	Used to generate symptoms that correspond with illness or disease.	https://drive.google.com	http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3635219/
Google Trends	Public web facility by Google that shows how often a particular search-term is entered relative to the total search-volume across various regions of the world, and in various languages.	Used to identify trends geographically, using the keywords from Many Eyes, Google Scholar and PubMed.	http://www.google.com/trends/	http://en.wikipedia.org/wiki/Google_Trends If Wikipedia will not suffice, I can find another source that describes it. (ie. Nature paper that used it to quantify trading behavior.)
Many Eyes	Free site from IBM that provides data visualization tools. Site allows users to upload datasets and produce graphic representations.	Used to visualize relationships in abstracts and help formulate query or key words.	http://www.many-eyes.com	http://www.many-eyes.com (Google search)
Nice Translator	Site that provides an improved interface for translating text	Used to translate terms (symptoms) and identify other languages within	http://www.nicetranslator.com/	http://nicetranslator.com/blog/about

	on the Web.	a text.		
Patients Like Me	A health data-sharing platform.	Used to identify self described symptoms and treatments.	http://www.patientslikeme.com/	http://www.patientslikeme.com/about
PubMed	Provides free access to the MEDLINE database, of indexed citations and abstracts to medical, nursing, dental, veterinary, health care, and preclinical science journal articles.	Used to search for literature, in order to get an idea of factor and information landscape of disease in question.	http://www.ncbi.nlm.nih.gov/pubmed	http://www.nlm.nih.gov/services/pubmed.html
Tweet Archivist	*Twitter* analytics tool used to search, archive, analyze, visualize, save and export tweets based on a search term or hashtag.	Used to capture geocoded tweets.	http://www.tweetarchivist.com/	http://www.tweetarchivist.com/ (Google search)

Table 3. Benefits and disadvantages of passive web-based methods for exploring disease factors (the look, listen, and learn paradigm).

Benefits (Pros (+))	Disadvantages (Cons (-))
Extremely useful and relevant tool for **"real-time" screening** of a wide-cast set of variables and undiscovered factors with minimum burden	**Not structured** epidemiological queries
Passive interrogation methods reduce Hawthorne effect and study participant biases common to traditional active survey methods	**"Web-savvy" demographics may vary** from actual population demographics (age, gender, ethnicity, and socioeconomic status)
Can be modified on the fly by language (i.e. use nicetranslator.com), terms, rapidly re-generated, and is essentially "**free**" in terms of cost	Social media monitoring by some (for instance government) may **appear "big brother"**...what are the ethical implications?
This form of information modeling can give rise to **understanding either data gaps, population variability, or technology penetration**	**Technology penetrance** in certain populations may be an issue (some counties, states, and countries have fewer technology resources for a variety of reasons, including socioeconomic factors, government censorship, etc.