UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

Online behavioral patterns in a health crisis setting: the 2009 pandemic.

Cláudio Haupt Vieira

DISSERTAÇÃO
MESTRADO EM BIOINFORMÁTICA E BIOLOGIA COMPUTACIONAL
ESPECIALIDADE EM BIOLOGIA COMPUTACIONAL

Dissertação orientada por:
Doutora Joana Gonçalves de Sá
Prof.ª Dr.ª Lisete Maria Ribeiro de Sousa

2017

Resumo

As doenças infecciosas representam um risco permanente para a humanidade. Existem duas abordagens naturais para tentar mitigar este risco: por um lado é necessário diminuir a probabilidade de infecção; por outro, tentar minimizar o seu impacto, caso o surto aconteça. Neste sentido, as instituições de saúde pública desenvolveram, ao longo de muitos anos, sistemas de monitorização que possam permitir uma deteção atempada e fiável de diferentes doenças e surtos epidémicos. Uma deteção em fases iniciais de um surto permite colocar em prática medidas que visam à contenção do mesmo, impedindo a sua progressão.

O caso da gripe é particularmente interessante. A gripe é uma doença infecciosa sazonal e todos os anos origina surtos epidémicos, durante a época fria de cada hemisfério. Estes surtos de gripe sazonal causam até 500 000 mortes anualmente e sobrecarregam os sistemas de saúde, suscitando a necessidade de constante monitorização. Estes sistemas de monitorização de gripe, desenvolvidos e apoiados por instituições de saúde pública como o CDC dos Estados Unidos ou o Europeu ECDC, publicam relatórios semanais com informação relevante para acompanhamento e prevenção de surtos. Estes incluem estatísticas de gripe, comparação com anos anteriores e estimativas da prevalência na população. Os dados na origem destes relatórios são produzidos por uma rede de médicos (Médicos Sentinela). Estes oferecem dois tipos de informação: primeiro, reportam sobre o número de doentes que se deslocam a unidades de sáude primárias com sintomas de gripe, permitindo uma estimativa da prevalência, e segundo, recolhem e enviam para análise amostras de doentes com sintomas de síndrome gripal, que permitem não só validação do diagnóstico, mas também a identificação de estirpes circulantes. Este sistema é, muito possivelmente, dos mais eficientes do mundo, mas tem duas limitações principais: só utiliza informação de doentes que procurem serviçoes médicos, pelo que só uma pequena fração do verdadeiro número de pessoas com gripe é registado pelos sistemas de monitorização. Para além disso o processo é moroso e geralmente resulta em relatórios com um desfasamento típico de duas a quatro semanas. Este desfasamento implica que as decisões de saúde pública se baseiam em informação desatualizada, encurtando o período de atuação para a contenção de um eventual surto.

Na era digital a procura por informação ocorre frequentemente na Internet, nomeadamente através de motores de pesquisa como o Google, e a pesquisa por questões de saúde é um hábito cada vez mais frequente. Quando um indivíduo infetado com gripe pesquisa no Google, por exemplo por sintomas que apresente, deixa um vestígio desta atividade. Se vários indivíduos infetados com gripe pesquisam em simultâneo por sintomas gripais então a atividade coletiva destes indivíduos pode constituir um sinal representativo de atividade gripal, o que fornece uma alternativa logisticamente e economicamente mais apelativa que os métodos tradicionais.

Assim, o ideal de métodos de monitorização online ganhou tração com o lançamento

da plataforma *Google Flu Trends* (GFT). O GFT agregava pesquisas online relacionadas com a gripe para obter um sinal de atividade gripal em tempo real, oferencedo uma solução para os problemas associados com os métodos de monitorização tradicionais. No entanto, o GFT errou nas previsões da magnitude da primeira onda da gripe pandémica de 2009 e, mais tarde, na magnitude da gripe epidémica sasonal de 2013, levando ao abandono do projeto. As razões para ambas as falhas não são inteiramente conhecidas, mas sabe-se que tanto o evento pandémico de 2009 como o evento epidémico de 2013 - ambos severos - estiveram associados a padrões de pesquisa online irregulares e a uma grande cobertura mediática.

Na prática, o modelo do GFT não conseguiu distinguir pesquisas associadas à atividade gripal, de pesquisas associadas a outros fatores. Isto expôs uma limitação intrínseca dos modelos de monitorização online: uma infeção de gripe não é a única motivação, nem possivelmente a mais forte, para as pessoas se interessarem pelo evento e até expectável, dada a severidade de uma pandemia, que as pessoas desenvolvam diferentes graus de interesse, que podem variar entre curiosidade, medo, ou infeção de facto.

Se não são infeções de gripe que levam indivíduos a pesquisar por termos relacionados com gripe durante uma pandemia, então é fundamental conhecer a sua motivação. Nesta tese exploramos diferentes formas de melhorar a análise de dados *online*, sob a hipótese de que deve ser possível utilizar esta informação para melhorar a monitorização da reação o público a uma gripe pandémica. Se conseguirmos identificar quais as motivações que induzem os indivíduos a pesquisar por determinados termos e quais os fatores que modulam essas motivações, então podemos separar pesquisas motivadas por infeção de gripe de pesquisas motivadas por outros factores, permitindo um sinal mais preciso da atividade gripal.

A pandemia de 2009 fornece uma excelente oportunidade para testar as nossas hipóteses por duas razões, 1) porque ocorreu numa altura em que o uso de Internet já era prevalente e 2) porque a gripe pandémica foi extensivamente estudada nos contextos biológico, psicológico e sociológico. Estes estudos geraram uma grande diversidade e riqueza de dados que podemos utilizar. Ao nível do contexto biológico, os casos suspeitos de gripe pandémica foram testados em laboratório, originando curvas epidemiológicas precisas.

No contexto psicológico vários questionários foram realizados ao longo do período pandémico para compreender a reação do público à pandemia, uma vez que o comportamento do público é determinante na contenção de transmissão viral. No contexto sociológico foram realizadas várias análises da atuação dos media e mesmo das entidades de saúde pública. A atividade mediática relativamente à pandemia foi muito elevada em fases iniciais mas rapidamente decresceu para niveis baixos.

Extraímos então séries de pesquisas do Google da Alemanha (GT-DE) e dos Estados Unidos (GT-US) e do Wikipedia em inglês (Wiki-EN) ao longo do período pandémico. Tentámos cobrir o máximo possível de variação de termos relacionados com a gripe: sintomas, vacinação, antivirais, comportamentos de higiene, instituições de saúde pública, entre outros.

Como possíveis variáveis explanatórias de comportamentos online, extraímos contagens de notícias relacionadas com gripe ao longo do período pandémico. Para além das notícias extraímos também o número de casos de gripe pandémica confirmados em laboratório. Paralelamente estimámos os níveis de ansiedade e de perceção de risco do público através dos dados de 17 questionários realizados em 9 países diferentes ao longo do período pandémico.

Iniciámos a análise através de clustering hierárquico de modo a inferir como as diferentes séries de pesquisa se relacionaram. Apesar de todas as séries obtidas serem referentes à gripe observámos um comportamento díspar entre estas. O clustering hierárquico suportou esta observação, ao distinguir dois grupos principais. Posteriormente utilizámos análises de correlação de Pearson, regressões lineares e testes de causalidade de Granger, entre cada um dos grupos, com o número de notícias relacionadas com a gripe e com o número de casos de gripe. Descobrimos que um grupo de séries de pesquisa está mais associado à actividade dos media e que o outro grupo de séries de pesquisa está mais associado à actividade gripal. Posteriormente analisámos através da correlação de Pearson e de regressão linear a associação de cada grupo com os níveis de ansiedade e de percepção de risco. Descobrimos que o grupo associado à actividade dos media está mais associado aos níveis de ansiedade e que o grupo associado à actividade gripal está mais associado à percepção de risco.

Deste modo os nossos resultados indicam que é possível distinguir entre motivações e que estas levam a diferentes padrões de pesquisa. A nossa abordagem permitiu também identificar termos que demonstraram menos sensibilidade à actividade mediática e que se correlacionaram com o número de casos de gripe. Estes termos são menos passíveis de conterém ruido, oferencendo a possibilidade de um sinal mais preciso de previsão de actividade gripal.

Adicionalmente, e uma vez que comportamentos como ansiedade e percepção de risco estão a associados a diferentes séries de pesquisas, este sistema possibilita a monitorização da reacção do público durante o desenvolvimento da pandemia, informação muito útil para fins de sáude pública. Os nossos resultados também sugerem que os media tiveram um efeito preponderante na maioria das series de pesquisa, mesmo aquelas que representaram adequadamente o numero de casos de gripe, e que a monitorização da atenção mediática é fundamental quando se utilizam dados de comportamento *online* para estimar comportamentos *offline*.

Assim, pensamos que este trabalho mostra ser possível refinar a análise de dados para distinguir entre diferentes tipos de comportamentos online. Em termos práticos, este novo sistema tem um grande potencial para complementar sistemas actuais de monitorização, para além de revelar uma grande riqueza e diversidade de comportamentos.

Palavras Chave: prospeção de dados, comportamentos online, saúde pública, pandemia, gripe

Abstract

Seasonal flu places a heavy burden on both human populations and health care services every year, warranting permanent surveillance. Online-based surveillance models harness the collective online search activity of flu-infected individuals to provide real-time monitoring of flu activity. These models assume that most flu-related online behavior is motivated by a flu infection. However, when the flu pandemic emerged in 2009 it resulted in abnormal search behaviors that confounded these models, as several reasons, beyond infection, can motivate individuals to seek flu information. In practice, and despite their potential, current models cannot distinguish whether such activity is related with actual flu infection or not, rendering them useless, at least in pandemic settings.

If the different motives that prompt flu-related searches can be pinpointed, then this information can be used to train the models to recognize what is infection-motivated and what is not. Moreover, if online behaviors reflect real-life behaviors, then valuable public health insights can be extracted by analyzing the public's online response to a pandemic.

To test these assumptions, we collected flu-related online search trends regarding the pandemic period. We estimated real-life behaviors, anxiety and risk perception, through data obtained from surveys conducted during the pandemic. As possible explanatory variables of online search trends, we collected flu-related media coverage as well as laboratory-confirmed flu cases.

We found that a specific set of search trends was more associated with media activity, whereas another set of search trends was more associated with flu infections. The media-related search trends proxied the public's anxiety levels and the infection-related search trends proxied the public's risk perception.

Having determined which factors correlated with specific search trends, and what real-life behaviors might have corresponded to these search trends, our findings place online sources as suitable tools for monitoring the public's response to a flu pandemic. Our findings additionally support the possibility of separating search trends that are more sensitive to media activity and search trends that are more sensitive to flu activity. Thus, we provide proof-of-principle that it should be possible to infer human behaviour from online behaviour and, in practical terms, our system is flexible and general enough to be applied both to pandemic and seasonal flu, as well as to other infectious settings.

Keywords: data mining, online behavior, public health, pandemic, influenza

Acknowledgements

Obrigado Joana pela oportunidade! Obrigado pelas correções Lisete! Obrigado Paulo pela tua ajuda incansável!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Infectious diseases pose great health risks to human populations worldwide. To mitigate these risks, public health institutions have set up surveillance systems that attempt to rapidly and accurately detect disease outbreaks. An early outbreak detection, followed by a timely response can limit and even stop the outbreak.

Seasonal flu outbreaks are responsible for up to half a million deaths annually and place a significant burden on health care systems (Lozano *et al.*, 2012; Molinari *et al.*, 2007). Flu surveillance systems used by public health institutions, such as the CDC and ECDC, publish weekly reports that seek to anticipate the onset of an influenza outbreak. These systems rely on small networks of health professionals that register and collect samples from influenza-like illness (ILI) occurrences during clinical visits. Despite its many advantages, this system has two main problems: first, as not all individuals seek health care when they experience flu symptoms, many flu cases tend to go unreported (Peppa *et al.*, 2017); second, the whole process is slow, usually resulting in reports that have a typical lag of two weeks (Won *et al.*, 2017). Lagged reports imply decisions based on old information, which is far from optimal as it curtails the window of opportunity to effectively respond to an outbreak.

In the digital era, it should be possible to use online and large scale information to support the decision-making process, in a timely and cost effective way. Online behaviour, such as searches on Google or Wikipedia, might prove to be very relevant tools, as health-seeking is a prevalent habit of online users (Fox, 2006) and their collective search activity has been proposed as a possible source of real-time indirect measures of ILI (Eysenbach, 2002). In fact, and in the specific case of influenza, there have been several reports matching online activity with "real world" epidemics. Hickmann *et al.* (2014); Lamb *et al.* (2013); Sharpe *et al.* (2016); Won *et al.* (2017) have shown that the collective search activity of flu-infected individuals, seeking health information online, provides a representative signal of flu activity in real-time without the need of clinical visits. And the potential of online-based surveillance methods gained large support with the launch of Google Flu Trends (GFT) in 2008. GFT attempted to predict the timing and magnitude of influenza activity by aggregating flu-related search trends and, contrary to traditional surveillance methods, GFT provided reports in near real-time (Ginsberg *et al.*, 2009). It seemed like a solution that would fill the gaps of traditional surveillance methods. However, GFT's notoriety was short-lived: despite being trained with large amounts of flu-related search trends, the GFT model underestimated the magnitude of the nonseasonal 2009 pandemic and

overestimated the magnitude of the severe seasonal flu outbreak in 2013, in the US. The GFT authors suggested that high media activity on the 2013 flu outbreak possibly led to abnormal flu-related Google search trends, possibly leading to GFT's overestimation in that year (Copeland *et al.*, 2013). Likewise, Cook *et al.* (2011) described abnormal Google flu-related search trends in the US during the nonseasonal 2009 pandemic.

These GFT's failures re-enforced existing skepticism over online-based analysis as possibly effective surveillance systems: at least in these two instances, GFT's algorithms could not distinguish between online behaviour guided by flu infection and behaviour guided by other factors (Lazer *et al.*, 2014). This made clear that a flu infection is not the sole (and perhaps not even the strongest) motivation for individuals to seek flu-related information online, particularly during extraordinary flu phenomena, such as a flu pandemic. Indeed, it is reasonable to expect individuals to have various degrees of interest in a flu pandemic, ranging from curiosity to fear, to actual disease. However, there is no *a priori* reason for this diversity in motivations to be seen as a limitation instead of as a possible asset: with more research and better designed algorithms, this richness in online behaviour might help us, one day, not only to better track diseases, but also to deepen our knowledge of human behaviour(s) in a quantitative and systematic way that never existed before. In fact, and despite the described limitations, there are several successful examples of using online behaviour as proxies for "real-world" behaviour in disease settings and many others areas (Choi & Varian, 2012; Moat *et al.*, 2014; Stephens-Davidowitz, 2014; Vosen & Schmidt, 2011; Won *et al.*, 2017).

Here, we argue that the 2009 flu pandemic provides an overall excellent opportunity to study the diversity in motivations, hidden behind apparently similar online behaviours, as the pandemic was extensively researched from the biological, psychological and sociological perspectives. First, precise signals of pandemic flu infections were obtained through large-scale laboratory confirmations (Panning *et al.*, 2009). Second, a vast number of surveys were conducted throughout the pandemic (Nguyen *et al.*, 2011; Tooher *et al.*, 2013), gauging emotional and psychological factors such as anxiety levels and perceived risk. Third, several studies analyzed the media's behavior during the pandemic (Duncan, 2009; Klemm *et al.*, 2014; Reintjes *et al.*, 2016), including the collection of news pieces and news counts. Fourth, and importantly, the pandemic emerged at a period marked by widespread Internet usage (Seybert & Lööf, 2010) and several online datasets have been made available (including the collective behavior of millions of users through their search trends on Google and Wikipedia).

Therefore, and by aggregating and comparing available data from these varied sources, we expect to uncover underlying insights into the public's online response to the pandemic and answer to main questions: 1) Can we use online behaviour as a reliable proxy for offline behaviour, in a flu pandemic setting and 2) can we distinguish between online behaviour driven by infection and driven by other factors, particularly flu-related media activity. b In the following sub-sections we start by providing a brief overview of the 2009 flu pandemic and the current surveillance mechanisms. We add some background on media coverage of crisis and present some context as to why understanding human behaviour in a risk-prone setting is fundamental, from the public health stand point. In the Methods section we describe the datasets and mathematical and computational tools that we used in the analysis. We present our comparisons in the Results, to show that the behaviours can indeed be distinguished. Finally, we discuss the strengths and limitations of this study from the epidemiological and computational perspectives.

1.1.1 Case study: the 2009 flu pandemic

The pandemic Influenza A(H1N1)09pdm strain (pH1N1) emerged in Mexico in February 2009 (Mena *et al.*, 2016) and was later confirmed to contain a unique genetic combination of both North American and Eurasian swine influenza lineages that had never circulated in humans before (Garten & Davis, 2009). By June 2009, pH1N1 had spread globally with around 30 000 confirmed cases in 74 countries. This prompted the World Health Organization (WHO) to declare the 2009 influenza pandemic - the first of the 21st century. In most countries pH1N1 displayed a bi-phasic activity: a spring-summer wave and a fall-winter wave (Brammer *et al.*, 2011; Devaux *et al.*, 2010). The fall-winter wave was overall more severe than the spring-summer wave as it coincided with the flu season (in the Northern Hemisphere), which provided optimal conditions for flu transmission (Shaman & Kohn, 2009). The pandemic was officially declared to be over in August 2010 and by then 214 countries had reported laboratory-confirmed pH1N1 cases. A total of 18 449 laboratory-confirmed pH1N1 attributable deaths were counted (WHO, 2009), but more recent studies argue that pH1N1 associated mortality was 15 times higher than the official number (Dawood *et al.*, 2012). Simonsen *et al.* (2013) states that even considering a corrected mortality estimate, the pandemic was overall less severe than an average seasonal flu. Regardless of its mildness, the pandemic flu still infected many and led to an overall increased awareness. In part because of the general concern, during the pandemic most suspected pH1N1 infections were tested in laboratory, resulting in very complete epidemiological curves. These can be used as our ground truth for the actual number of cases.

Media coverage and Risk Communication

Risk communication is a cornerstone of pandemic management as it conveys important information to the public regarding health behaviors that might mitigate flu transmission (Jefferson *et al.*, 2008). Public health institutions take advantage of the media's wide outreach to transmit information to the public. However, this reliance on media poses a conflict of interest, as media emphasizes on threat whereas public health officials emphasize normalcy (Anzur, 2000). Media's coverage often goes beyond transmitting public health announcements, as it inevitably selects and amplifies certain events based on their news-value, usually conflicting or dramatic, thereby shaping the the public's social construction of a crisis (Ma, 2005; Singer & Endreny, 1993). While the pandemic was overall mild, its severity was initially unknown. The uncertain severity combined with the rapid succession of events, that started from localized outbreak to a widespread transmission in North America, led the WHO to declare an international health crisis in late April 2009. This evoked an abnormally highly media activity (Duncan, 2009; Smith *et al.*, 2013) that quickly subsided, sustaining only low levels of activity across the remaining pandemic period. Regardless of the content or tone of news items, the initial sheer volume of news was enough to cause public alarm (Klemm *et al.*, 2014). Considering the media's prominent role in the public's perception of the pandemic crisis, we used media activity (in terms of flu-related news counts) as a possible explanatory variable of online search trends.

Public's response

Given lack of initial knowledge regarding the severity of the pandemic, the large (and sometimes confusing) amounts of information offered by public health officials, and the high and somewhat exaggerated media coverage, it should not be surprising to find changes in the public's anxiety

levels and perceived risk. The public's response is expected to change over the development of a pandemic, adjusting to novel information. Given the crucial role that the public plays in containing or spreading flu (Fenichel *et al.*, 2011), monitoring the public's response allows public health official to assess whether the employed policies are being effective and if not, why. Surveys produce empirical data that can used to increase knowledge on the public's response to specific phenomena (Kelley *et al.*, 2003) and several surveys were conducted throughout the pandemic period, asking a large number of respondents to report on their worries over the pandemic or if they thought they were at risk of contracting the pandemic flu. These surveys provide a vast source of empirical data on real-life behaviors, that can be used to gain insight into real-life concerns and ask whether they are reflected on online behaviours. Anxiety arguably measures an emotional response to a threat, whereas risk perception measures a cognitive dimension of the risk posed by the threat (Sjöberg, 1998). Therefore, we estimated levels of risk perception and anxiety from several surveys, as these behaviours were found to be substantial determinants of the public's compliance with health directives during the pandemic (Chan *et al.*, 2014; Gaygısız *et al.*, 2012; Jones & Salathé, 2009; Prati *et al.*, 2011; Rubin *et al.*, 2009; Taylor *et al.*, 2009; Walterdrkide *et al.*, 2012). Moreover, since one offers a more emotional perspective and the other a more rational one, we were curious as to how they would correlate with our other datasets.

Chapter 2

Methods

We extracted weekly data from Google search trend from Germany (GT-DE) and United States (GT-US) and the English Wikipedia (Wiki-EN) article-views. We collected search trends that cover various aspects of a pandemic flu, such as flu symptoms, vaccination, antivirals, hygiene behaviors, institutions and flu pandemic circumstantial terms. While we were not able to collect the exact search trends for the three datasets, there are comparable overlapping search trends. In total, 49 search trends were collected for GT-US, 31 for GT-DE and 25 for Wiki-EN. All collected search trends are shown in page 61.

As possible explanatory variables, we extracted weekly influenza-related news counts (media activity) and the weekly laboratory-confirmed pH1N1 cases (flu activity) in both countries. In parallel, we estimated the public's flu-related anxiety and risk perception based on data collected from 17 different surveys conducted in 9 different countries throughout the 2009 pandemic, covering a period of 10 months from April 2009 to January 2010. A summary of the extracted data from each survey is shown in pages 64 and 65.

We used clustering analysis to infer the similarity between the different search trends of each online dataset (GT-US, GT-DE, Wiki-EN) and to possibly extract patterns. We then tested the search trends contained in each different cluster with media and flu activity using Pearson's correlation, Linear regression and the Granger Causality test. The estimated Anxiety and Risk perception were also tested with each search trends cluster, but only with Pearson's correlation and Linear regression – Granger causality test was not applied in this data due to insufficient data points.

Table 2.1: Overview of the collected datasets.
Anxiety and Risk perception summary is shown in pages 64 and 65

	Media activity	Flu activity	Search trends
United States	Online Print TV newscast	pH1N1 lab-confirmed CDC-NREVSS (Flahault *et al.*, 1998)	GT-US Wiki-EN
Germany	Print TV newscast (Reintjes *et al.*, 2016)	pH1N1 lab-confirmed Robert Koch Institute (Krause, 2010)	GT-DE

2.1 Data Collection

2.1.1 Country Selection

Although we have collected data from Google Trends on a larger number of countries, we were not able to gather complete datasets for all of them. We only have full datasets for the United States and Germany (except Wikipedia). These are the countries we focus in the context of this thesis. We test our hypothesis in a global setting by analyzing these two distant countries in parallel.

2.1.2 Search trends

Google

Google provides an index of search activity of specific queries through the *Google Trends* (GT) API. This index measures the total number of searches for a particular query normalized by the total search volume in the specified geographical region and within the given time range (Stephens-Davidowitz & Varian, 2014). It is scaled by the maximum value, *i.e.* measured relative to the highest search point in the specified time. We extracted weekly and monthly data from July 2008-February 2009 for the prepandemic period, from March 2009 - July 2010 for the pandemic period, and from August 2010-August 2011 for the postpandemic period. We collected 49 search trends in the United States Google Trends (GT-US). For the German Google Trends (GT-DE) we translated each search trend used in GT-US to German, however, some search trends were not retrievable due to low search volumes, so only 31 were collected. Both datasets share a set of 31 search trends in common. Some search trends were also retrievable in GT-DE but not GT-US. When more than one search query is extracted through Google Trends API, the search trends are normalized such that a more widely searched term might push a less searched term to low SVI, possibly even 0. To avoid this we extracted each term's search trends at a time Stephens-Davidowitz & Varian (2014). While we lose information on the relative magnitude difference between queries, the trends over time are still meaningful.

Wikipedia

We extracted Wikipedia's article views through the unofficial API (`http://stats.grok.se`). Since Wikipedia does not provide information on the geographical origin of each page view, this makes the task of comparing Wikipedia's data to a specific country difficult. However, the largest proportion of the English Wikipedia's article views comes from the United States (43%) (Wikipedia, 2017) and has been successfully used for influenza onset prediction in the United States (Hickmann *et al.*, 2014; McIver & Brownstein, 2014). We could not collect data from the German Wikipedia, as the API was down as of June 2017.

As per Google, we extracted weekly and monthly page-views of several influenza-related articles, from July 2008-February 2009 for the prepandemic period, from March 2009 - July 2010 for the pandemic period, and from August 2010-August 2011 for the postpandemic period. However, as Wikipedia's articles designations are predefined we could not replicate all nor the exact search terms used in the Google analysis. We extracted 18 comparable keywords with GT-US. There are known sparsely distributed missing data points in this dataset due to Wikipedia's servers being down. We imputed these missing data points through an autoregressive integrated

moving average (ARIMA) process, which is specifically suited for time series data, implemented in the *imputeTS* R package (Moritz & Bartz-Beielstein, 2017).

2.1.3 Flu activity

We collected weekly counts of laboratory-confirmed pH1N1 cases in the United States and Germany. A laboratory confirmed case is defined as any tested case with positive detection of pH1N1 nucleic acids by RT-PCR (Panning *et al.*, 2009). We collected publicly available weekly counts of laboratory-confirmed pH1N1 cases from March 2009 to July 2010 in the United States, registered by CDC's National Respiratory and Enteric Virus Surveillance System (Flahault *et al.*, 1998). The data on weekly laboratory-confirmed pH1N1 cases from March 2009 to April 2010 in Germany was assessed by the *Robert Koch-Institut* and collected from the study by Krause (2010).

2.1.4 Media activity

For the *media activity* dataset, we collected news counts from both television, online and print sources in the United States. We collected TV news broadcasts from NBC, CBS, CNN, FOX and MSNBC networks containing the term "flu" or "influenza", from March 2009 to August 2010 through the Vanderbilt Television News Archive (`https://tvnews.vanderbilt.edu/`). For the online and print newspapers, we used the New York Times API to collect all of print and online NYT news containing the terms 'flu' or 'influenza'. As the weekly counts of the both NYT (online, print) and TV news broadcasts were highly correlated, we used the sum of both datasets in all the subsequent analysis. For the German media activity dataset we used data previously collected by Reintjes *et al.* (2016). These authors collected weekly influenza-related news counts from *ARD Tagesschau* (TV newscast), *Frankfurter Allgemeine Zeitung* (daily newspaper), *Bild* (tabloid) and *Spiegel* (weekly newspaper). This dataset spans the period from April 2009 to April 2010.

2.1.5 Surveys

In order to obtain an estimate on the public's risk perception and anxiety levels regarding pH1N1 over the course of the pandemic event, we collected data from 18 surveys conducted in 9 different countries, covering the period between April 2009 to January 2010. We considered surveys containing questions regarding *concern, worry, anxiety* for the *"anxiety"* dataset and *likelihood, risk* or *susceptibility* to a pH1N1 infection for the *"Risk perception"* dataset. A summary of the articles, periods covered and the extracted value is shown in Appendix B.2.

Different studies used different metrics and response scales making comparisons difficult (Moeller, 2015). In order to obtain comparable values we normalized the surveys in two ways: when the percentage or proportion of respondents in each option was known, we collected the proportion of respondents that displayed intermediate to high levels of risk perception and anxiety, regardless of the scale (eg. if the survey asked participants how anxious they were, regarding the pandemic, on a scale from 1 to 5, with 1 being very little, and 5 very much, we collected the proportion that answered 3 or more); Some articles (eg.(Rudisill, 2013)) did not provide such proportions, providing the average obtained from Likert scales. In these cases we considered the averaged Likert value over the maximum value of the Likert scale, as described in Little (2013).

This normalization (to the maximum) allowed us to build a monthly estimate of risk perception and anxiety levels by computing the mean and standard deviation for each time point. As none of the collected surveys regarding anxiety levels covered October 2009 this value was linearly interpolated using the *NumPy* Python library (Walt *et al.*, 2011).

.

2.2 Dataset analysis and comparison

We compared the different datasets through correlation, regression and causality analysis. In addition, hierarchical and fuzzy clustering were used to extract patterns from the online search trends data.

2.2.1 Correlation analysis

To test if the collected datasets vary linearly, we used the Pearson's correlation coefficient (Pearson, 1901). It measures the strength of a linear relationship between two continuous variables. It can be calculated as the sum of products of deviations of the two variables divided by the square root of the product of the two sums of squares:

$$R_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{2.1}$$

where X and Y represent two different random variables; and both $X_1, X_2, ..., X_n$ and $Y_1, Y_2, ..., Y_n$ correspond to the sampled populations from X and Y, respectively. Also, $\bar{X}$ and $\bar{Y}$ correspond to the sample means of X and Y respectively.

Pearson's correlation coefficient varies from -1 (a perfect negative correlation) to +1 (a perfect positive correlation), with a value of 0 indicating no linear relationship at all. For a Pearson correlation variables should be continuous. For instance, Google provides normalized data in a discrete 0-100 scale, where Wikipedia provides absolute counts of page-views. The collected datasets are, therefore, approximately continuous.

To reduce the possibility that the observed Pearson's correlations occurred by chance, we employed the correlation t-test to establish statistical significance. For this analysis we considered a level of significance (α) of 0.05, with the null hypothesis being no correlation between the two variables. In addition we also used a two-sample t-test (Snedecor & Cochran, 1989) to determine if mean correlations are equal in different groups. We considered a level of significance (α) of 0.05, with a null hypothesis H_0 of equal means.

2.2.2 Linear regression

Regression analysis is used for explaining or modeling the relationship between a single variable Y, called the *response variable*, and one or multiple *explanatory variables*. For our analysis we used a simple linear regression, which models the relationship between the response variable and one explanatory variable. The best-fitting line is obtained by finding the linear regression parameters that minimize the residual sum of squares (RSS), which can be estimated through least-squares method (Iksoon, 1996). In order to assess the regression goodness of fit we computed the coefficient of determination, R^2 , which describes the proportion of variation explained by

the fit. R^2 values vary from 0 to 1, with 1 indicating a perfect fit and 0 indicating that none of the variation in the data is explained by the fit. For a simple linear regression, the R^2 is equivalent to the squared Pearson's correlation.

To check if the linear regression assumption of normality is not violated, we used the Shapiro-Wilko test of normality (Shapiro & Wilk, 1965) with the R function *shapiro.test*. For this analysis we considered a level of significance (α) of 0.05 with the null hypothesis (H_0) of the data being normally distributed.

A t-test was applied to determine the statistical significance of the simple linear regression fit. For this analysis we considered a level of significance (α) of 0.05 with a null hypothesis (H_0) being no linear association. The R *lm* function was used to estimate the linear regression model parameters and respective hypothesis testing.

2.2.3 Causality Analysis

We used Pearson's correlations to test the association between online data, media activity, pH1N1 activity, anxiety and risk perception. However, time series data is usually dependent on time and Pearson's correlation is more appropriate for independent variables, so it can possibly provide misleading statistical evidence of a linear relationship *i.e. spurious correlation*. To rule out the possibility of spurious correlations we employed in parallel a more sensitive test, Granger Causality. A statistically significant Granger-causality result in line with statistically significant correlations provide an indication of non-spuriousness, adding robustness to the results.

2.2.3.1 Granger Causality Test

To gauge causality, we used the Granger causality test (Granger, 1969), that assumes that if an event A precedes an event B, then it is possible that A is causing B. The Granger causality states that $x_i(t)$ time series has a causal link to another time series $x_j(t)$, if the lagged values of x_i can predict the values of x_j. The Granger causality score from the variable i to the variable j ($i \to j$) is computed as in the following steps. First, an autoregressive model (i.e. where the output variable depends linearly on its own previous values and on a stochastic term) of order L is fitted to the $x_j(t)$ time series. The x_j time series is thus expressed as as:

$$x_j(t) = \sum_{\tau=1}^{L} a_\tau \cdot x_j(t-\tau) + \epsilon_j(t), \tag{2.2}$$

where a_τ is a matrix with the fitted model parameters for every τ and ϵ_j a vector of the residuals. Secondly, a bivariate autoregressive model is fitted on the x_j time series using the lagged values of x_i:

$$x_j(t) = \sum_{\tau=1}^{L} a_\tau \cdot x_j(t-\tau) + \sum_{\tau=1}^{L} b_\tau \cdot x_i(t-\tau) + \epsilon_{j|i}(t) \tag{2.3}$$

where a_τ, b_τ are the fitted model parameters and $\epsilon_{j|i}$ is the residual with variance $\sigma_{j|i} = \mathrm{Var}(\epsilon_{j|i})$. The Granger causality score $G_{i\to j}$ is thus defined by:

$$G_{i\to j} = \log\left(\frac{\sigma_j}{\sigma_{j|i}}\right). \tag{2.4}$$

The variance of the residuals from the bivariate autoregression model (predicted from lagged values of x_i and x_j) is compared with the variance of the residuals from the univariate autoregressive model (predicted from lagged values of x_j alone). The logarithm of the ratio of the variances of residuals is χ^2-distributed thus allowing to establish a statistical significance for each test. We considered a level of significance (α) of 0.05 for this analysis, with the null hypothesis H_0 being non-Granger causality. Higher or lower values of $G_{i\to j}$ indicate that the lagged x_i time series is accordingly more or less successful in predicting the future values of x_j under the autoregressive assumption. The order of the autoregressive model, considering a maximum of 4 lags, was selected based for each pair of series on the Aikaike's Information Criteria (AIC), a objective method that selects the best approximating model. We used the Granger-causality test implemented in the R *vars* package (Pfaff, 2008).

The Granger causality test assumes series stationarity *i.e.* statistical properties do not vary across time. In the case of non-stationarity the asymptotic distribution of the test statistic may not be valid under the null hypothesis. Before proceeding with the Granger-causality test we ensured all series were stationary to fit the requirements of the test. To fulfill the stationary requirement we computed the first differentiation of each non-stationary series to remove the trend. The first difference of a time series measures the series of changes from one period to the next, which can be described by the following equation:

$$Y_t^{'} = Y_t - Y_{t-1}. \tag{2.5}$$

Note that as any other causality or correlation test, the Granger causality test is not suited to test causal relationships in the strict sense, as we cannot exclude the possibility of spurious causality *i.e.* a *post hoc* fallacy. This test can however provide evidence in support of a hypothesis about causal links. We therefore refer to the results of this test as 'x_i *Granger-causes* x_j' or 'x_i does not *Granger-cause* x_j'.

2.2.4 Cluster Analysis

We collected the Google and Wikipedia searches on a large number of influenza-related terms. These terms and page views display different patterns, varying in time. In order to understand how these terms grouped and if they described any meaningful pattern we used cluster analysis. Cluster analysis is the task of dividing a set of objects into *a priori* unknown groups, or clusters. The definition of cluster varies with the employed technique but it usually implies a group of objects more similar (in a given measure) within their group and more dissimilar between objects of other groups. We used two different clustering methods, with similar, re-enforcing, results. First, we used hierarchical clustering because 1) dendrograms are informative visual representations of the arrangement of the clusters and 2) different numbers of clusters can be obtained by cutting a dendrogram at different levels.

However, hierarchical clustering is sensitive to noise and outliers and, due do its agglomerative or divisive algorithm, once a cluster is defined there cannot be a change of membership. Therefore, we also grouped our data using fuzzy clustering (See: 2.2.4.2), that offers a less rigid measure of cluster membership.

2.2.4.1 Hierarchical Clustering

Hierarchical clustering is a method that builds a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types, agglomerative and divisive. Agglomerative hierarchical clustering builds a hierarchy from individual objects by progressively merging clusters in a bottom-up fashion, whereas divisive hierarchical clustering works the opposite way, with all initial objects belonging to one cluster that is progressively subdivided in a top-down approach. We applied the agglomerative algorithm to our data using R *TSClust* library (Montero *et al.*, 2014). In order to decide which objects should be combined at each step of the agglomerative hierarchical clustering algorithm, a dissimilarity matrix containing the pairwise distance between all objects is required. A dissimilarity matrix is obtained through measuring the pairwise distance between all objects.

Time series clustering is not a trivial task. Dissimilarities conventionally used in clustering routines may not work adequately with time dependent data because they these do not take time inter-dependence relationship between values into account. However, since we want compare profiles between time series of equal length, then conventional distances should suffice (Montero *et al.*, 2014). We tested the Euclidean distance and a Pearson's correlation based distance (simply defined as $1 - \rho$) and both dissimilarities rendered relatively similar results. As we are dealing with scale-dependent data, and Pearson's correlation are scale-independent, we opted for the Euclidean measure as a dissimilarity metric between objects in clustering routines. The Euclidean distance between each time series x and y is defined as

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 \tag{2.6}$$

where n denotes the number of time points in each series. We applied the Euclidean distance in each clustering routine to the collected online series data (GT and Wikipedia), which were pre-processed to zero mean and unit variance using z-score normalization (using *R scale* function).

At each agglomerative step, as objects are clustered, the dissimilarity matrix needs to be updated with the new distances according to a linkage criterion. We used the Ward's linkage criterion (Ward Jr, 1963), as this method tends to build homogeneous and equal-sized hierarchies by minimizing the within-cluster variance. It assumes that a cluster is represented by its centroid, m, and measures the proximity between two clusters in terms of the increase of sum of the squared error (SSE), as described below.

$$d_{A,B} = \sum_{i \in A \cup B} ||x_i - m_{A \cup B}||^2 - \sum_{i \in A} ||x_i - m_A||^2 - \sum_{i \in B} ||x_i - m_B||^2 \tag{2.7}$$

The proximity between two clusters, A and B, is the magnitude by which the square of sums of their joint cluster is greater than the combined summed square in each of these two clusters. The combination of data points that yield the lowest sum of squares is chosen thus minimizing the total within-cluster variance. We also tested other linkage methods, such as average linkage, and all yielded consistent results.

Cluster Validation

Agglomerative hierarchical clustering algorithms do not require a predefined number of clusters, but we wanted to know how many clusters should be considered for subsequent analysis, *ie.* where

we draw the line on the dendrogram to separate the clusters. As we have no *a priori* knowledge of how the search trends are grouped, we cannot measure the accuracy attained by a clustering algorithm. We can, however, resort to internal validation methods, *i.e.* validation methods that measure the goodness of a clustering structure solely based on information contained in the data. These methods are based on two criteria: compactness and separation. Compactness can be measured through the within-cluster variance or distance. Separation measures how distinct clusters are between them. We selected three different criteria to evaluate the optimal number of clusters. We opted for Dunn's index (Dunn, 1974), Silhouette index (Rousseeuw, 1987) and the modified Davies-Bouldin (DB*) index (Kim & Ramakrishna, 2005). The three indices were computed using *R* package *dtwclust* (Sardá-Espinosa, 2017).

The Silhouette index measures the similarity of an object to the centroid of its own cluster compared to other clusters centroids. Dunn's index measures the maximum distance in-between objects of clusters and minimum distance between clusters. The DB* index evaluates the ratio between intracluster similarity and inter-cluster differences and computing the average overall the clusters. In Silhouette and Dunn's index, higher values indicate a better partition, whereas in the DB* index lower values indicate a better partition. We chose the optimal number of clusters based on a quorum between the three indices. If quorum was not reached, prevalence was given to the DB* index as it is less sensitive to noise and is overall among the best performing indexes (Kim & Ramakrishna, 2005; Liu *et al.*, 2010).

Cluster comparison

To compare similarity between different hierarchical cluster structures (dendrograms) we use the *Fowlkes-Mallows Index* (Fowlkes & Mallows, 1983). It is measured as follows

$$FM = \sqrt{\frac{TP}{TP+FP} \cdot \frac{TP}{TP+FN}} \tag{2.8}$$

Where TP is the number of points that are present in the same cluster in both A1 and A2 structure. FP is the number of points that are present in the same cluster in $A1$, but not in $A2$. FN is the number of points that are present in the same cluster in $A2$ but not in $A1$. Finally, TN is the number of points that are in different clusters both in $A1$ and $A2$ structures. This index ranges from 0 to 1, where a higher value indicates more similarity between clusters and a lower value lesser similarity between clusters. The result was extracted from the maximum FM index considering up to 5 clusters. Considering a level of significance (α) of 0.05, we performed permutation tests of the FM-Index with the null hypothesis H_0 being "not-similar" clusters. This was performed using the *dendextend* R package (Galili, 2015).

2.2.4.2 Fuzzy clustering

Time series usually display dynamic behavior over time, which should be taken into account when considering a cluster analysis. In a given range a time series might belong to a certain cluster but it might be closer to other clusters across time. This implies that *hard clustering* approaches, such as hierarchical clustering, might miss the underlying structure of time series. This type of problem can be solved by fuzzy clustering, where each time series is not exclusively assigned to one cluster: it is allowed to belong to multiple clusters with varying degree membership. This

method enabled us to discern series that displayed dynamic behaviors belonging to multiple clusters from series that are fixed on only one cluster. It allows gradual memberships measured as probabilistic degrees between $\{0, 1\}$. We applied a fuzzy c-means clustering method (Bezdek *et al.*, 1984) in parallel with hierarchical clustering routines, using the same distance measure (Euclidean) and an equal number of clusters as determined by cluster validation measures. We considered a low cluster membership value if below 0.75.

Chapter 3

Results

Despite the all the search trends being we have collected being related to influenza, our results show different behaviors between these search trends (Figures 3.1 to 3.3). The overall search trends can be described in the following manner: 1) a prominent peak is observed around April/May 2009 *"early peak"*; 2) another prominent peak is observed around April/May 2009: *"early peak"*; 3) other smaller peaks are observed; 4) some series display both early and later peaks, in varying magnitudes, 5) later peaks appear to be country-specific; 6) early peaks are similar across the three datasets.

Two distinct online behavioral patterns are observed.

We wanted to infer the similarity between all the search trends during the pandemic period, and if any common patterns could be extracted. To achieve this we applied clustering analysis in GT-US (Figure 3.1), GT-DE (Figure 3.2) and Wiki-EN (Figure 3.3) pandemic period data. The dendrograms across the three datasets reveal two distinct patterns in the three datasets, as supported by cluster validation indices (Appendix B.2). Correlation matrices rendered relatively concordant clusters (pages 46 to 48). Cluster 1 (C1) search trends are characterized by the early and later peaks, but generally with a more prominent later peak. Cluster 2 (C2) search trends are characterized by a very prominent early peak and a slight later peak. C1 search trends are more variable than C2 search trends.

In order to extract meaningful representations *i.e. a pattern* from the search trends of each cluster, we computed the centroid as the mean of all series contained in each cluster (Figure 3.4). Extracted centroids are at least highly correlated (above 0.8) with all series within its respective cluster. C1 and C2 centroids match the previous description of C1 and C2 search trends. Country-specificity of C1 centroids centroids is evident. C2 centroids are very similar (Figure 3.6). We use the centroids in subsequent analysis, but we analyze the centroids and each clusters search trends in parallel as as cross-confirmation to check if deviant results are obtained.

We then assessed the quality of hierarchical clustering by using fuzzy clustering. Despite the evident cluster dichotomy obtained through hierarchical clustering, fuzzy clustering results reveal an amount of uncertainty in how some search trends are clustered (*"Cm"* in Figures 3.1 to 3.3). Search trends with low cluster membership (<0.75) display a dynamic behavior in-between the patterns described by C1 and C2 (Figure 3.5). Such search trends are not too distant yet still not close enough to be assigned to the opposite clusters. A third cluster including these series is not supported by the cluster validation indices as the best partition (Appendix B.2). Most low

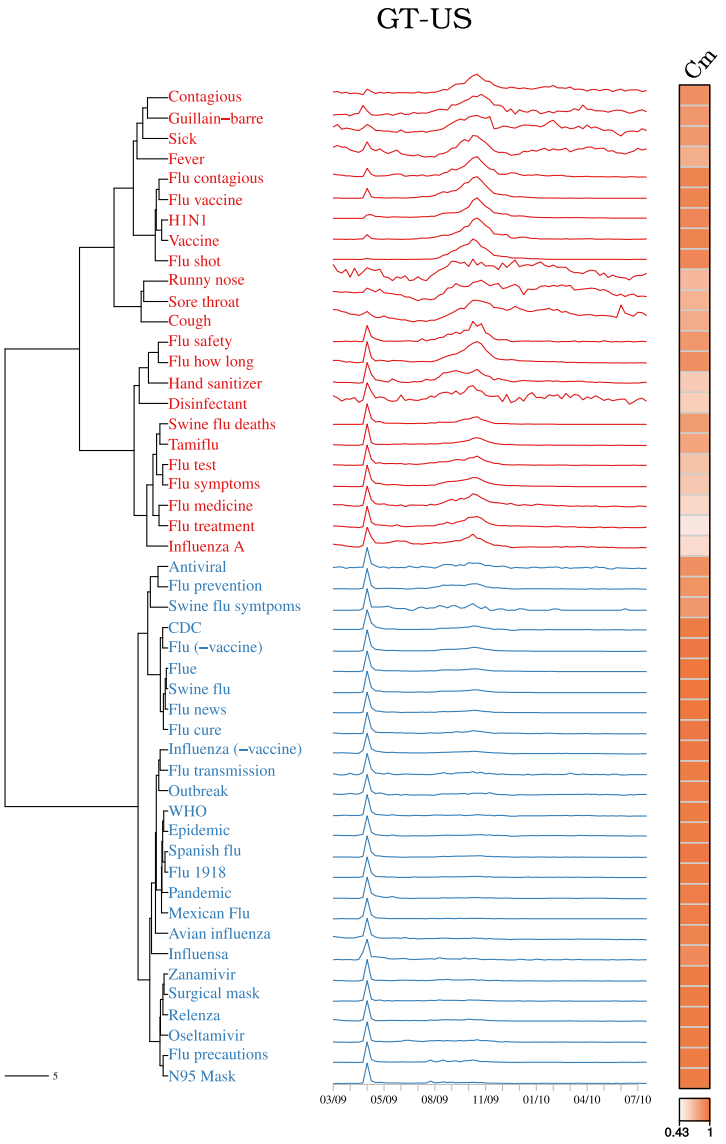

Figure 3.1: Wiki-EN dendrogram is shown in the left column. The dendrogram was obtained through agglomerative hierarchical clustering using Euclidean distance on z-score normalized data and Ward's linkage criterion. The period pandemic period (March 2009-July 2010) is used. Cluster validation indices support cutting the dendrogram in two clusters (Table T1), which are distinguished by the red color (Cluster 1) and blue color (Cluster 2). The middle column displays the time-series of each collected term. The results of fuzzy clustering (Cm) are shown in the right column, lower values indicate cluster membership uncertainty. Bottom left scale indicates the Euclidean distance.

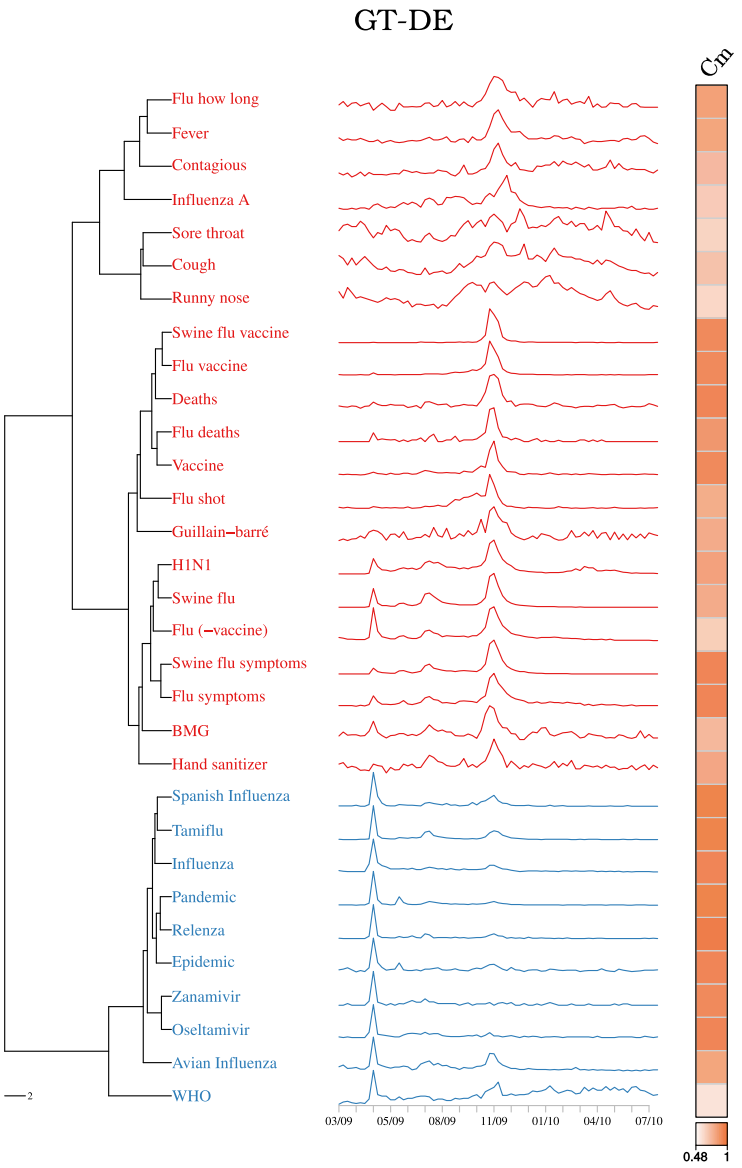

Figure 3.2: Wiki-EN dendrogram is shown in the left column. The dendrogram was obtained through agglomerative hierarchical clustering using Euclidean distance on z-score normalized data and Ward's linkage criterion. The period pandemic period (March 2009-July 2010) is used. Cluster validation indices support cutting the dendrogram in two clusters (Table T1), which are distinguished by the red color (Cluster 1) and blue color (Cluster 2). The middle column displays the time-series of each collected term. The results of fuzzy clustering (Cm) are shown in the right column, lower values indicate cluster membership uncertainty. Bottom left scale indicates the Euclidean distance.

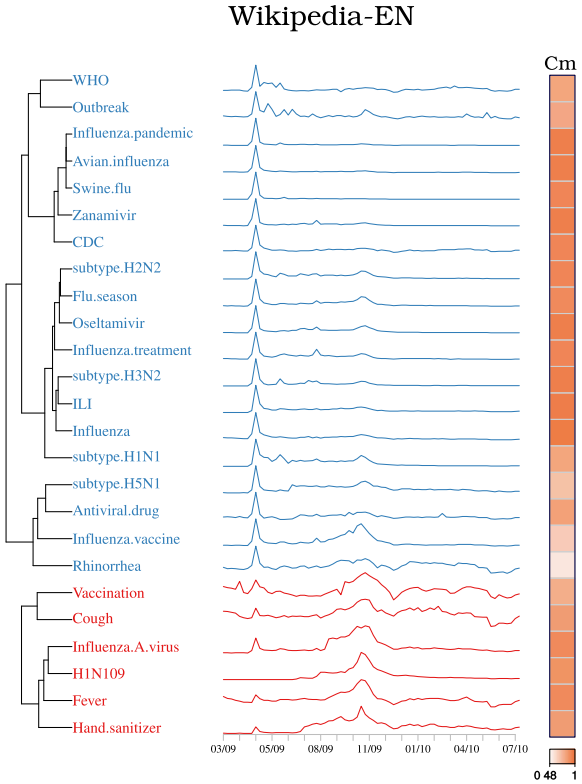

Figure 3.3: Wiki-EN dendrogram is shown in the left column. The dendrogram was obtained through agglomerative hierarchical clustering using Euclidean distance on z-score normalized data and Ward's linkage criterion. The period pandemic period (March 2009-July 2010) is used. Cluster validation indices support cutting the dendrogram in two clusters (Table T1), which are distinguished by the red color (Cluster 1) and blue color (Cluster 2). The middle column displays the time-series of each collected term. The results of fuzzy clustering (Cm) are shown in the right column, lower values indicate cluster membership uncertainty. Bottom left scale indicates the Euclidean distance.

cluster membership series belong to C1, which displays more variability than C2.

Search trends clusters differed in the peripandemic period.

In addition to the pandemic period, we extracted the peripandemic search trends for all three datasets to understand how these periods differed. The peripandemic and pandemic time series are shown in pages 43 to 45. Online behaviors during the pandemic period disturbed the seasonality of some search trends. Other search trends were specific to the pandemic period. In order to understand how the similarity between search trends varied before and after the pandemic, we applied hierarchical clustering to the prepandemic and postpandemic periods of GT-US, GT-DE and Wiki-EN. The prepandemic and postpandemic periods include seasonal influenza signal. We then tested the similarity of pandemic clusters with prepandemic and postpandemic clusters using the *Fowlkes-Mallows Index*, which varies between $\{0, 1\}$ with higher values indicating greater similarity. In all three datasets the pandemic and post-pandemic search trend clusters differed, despite maintaining a moderate and statistically significant similarity (Table 3.1).

Table 3.1: FM index between pandemic and peripandemic periods.
denotes $p < 0.05$

	Prepandemic	Postpandemic
GT-US	0.48*	0.53*
GT-DE	0.55*	0.52*
Wiki-EN	0.55*	0.53*

GT-US and Wiki-EN have highly similar search trends.

We wanted to infer the similarity between the overlapping search trends of the two different online platforms, Wiki-EN and GT-US. We computed the pairwise correlation between overlapping terms in GT-US and Wiki-EN ($n = 18$). Overlapping terms between the datasets were on average highly correlated (0.86 ± 0.12). In addition, GT-US and Wiki-EN C1 centroid. GT-US and Wiki-EN C2 centroids are very highly correlated (Figure 3.6). This suggests the search trends were very aligned during the pandemic period in both platforms, and that this similarity applies both to C1 and C2 search trends.

We then tested whether overlapping search trends high average correlation was specific to the pandemic period. The average pairwise correlations decreased to $0.53(\pm 0.19)$ in the prepandemic period and to 0.59 (± 0.26) in the postpandemic period. Therefore, both platforms were more aligned during the pandemic period than before and after.

C1 search trends are country-specific.

We tested the assumption of country-specific search trends by computing the average correlation between GT-US and GT-DE overlapping search trends during the pandemic period. GT-US and GT-DE overlapping search trends were on average moderately correlated 0.66 ± 0.26.

We then computed the average correlation between overlapping search trends of overlapping C1 and C2 search trends. GT-US and GT-DE C1 series are moderately correlated (0.57 ± 0.22) and both datasets C1 centroids are likewise moderately correlated ($R = 0.64$, $p < 0.05$).

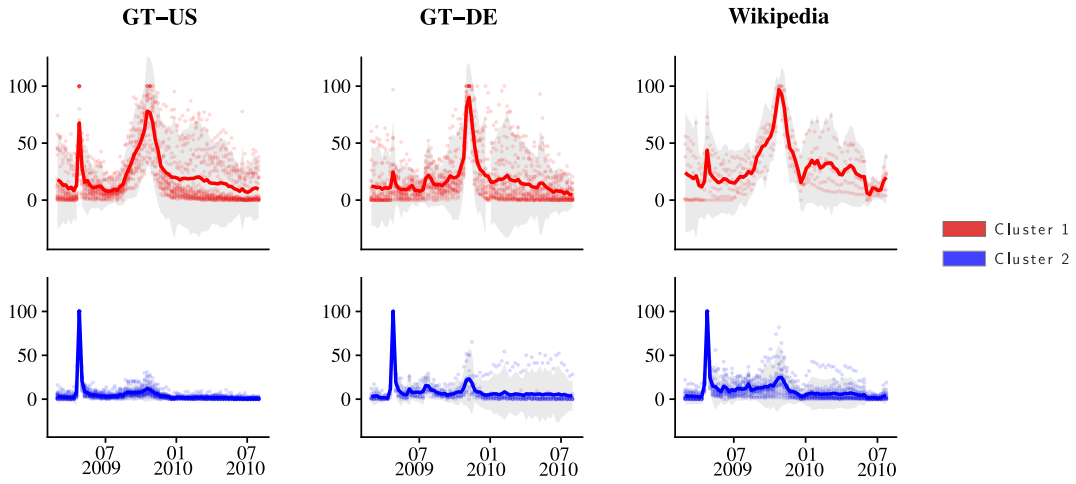

Figure 3.4: Centroid and standard deviation (grey shade) of Cluster 1 and Cluster 2 of Google Trends (US, DE) and Wikipedia. GT-US C1 $n = 26$, C2 $n = 23$; GT-DE C1 $n = 21$, C2 $n = 10$; Wiki-EN C1 $n = 6$, C2 $n = 19$.

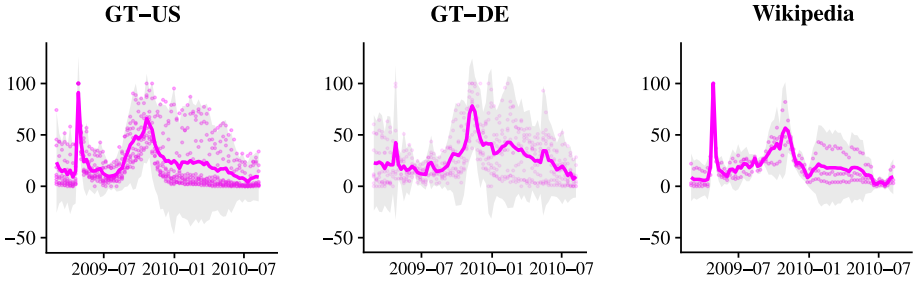

Figure 3.5: Centroid (mean) and standard deviation series of fuzzy objects of each dataset (<0.75 cluster membership certainty). In GT-US $n = 9$, GT-DE $n = 8$, Wiki-EN $n = 3$.

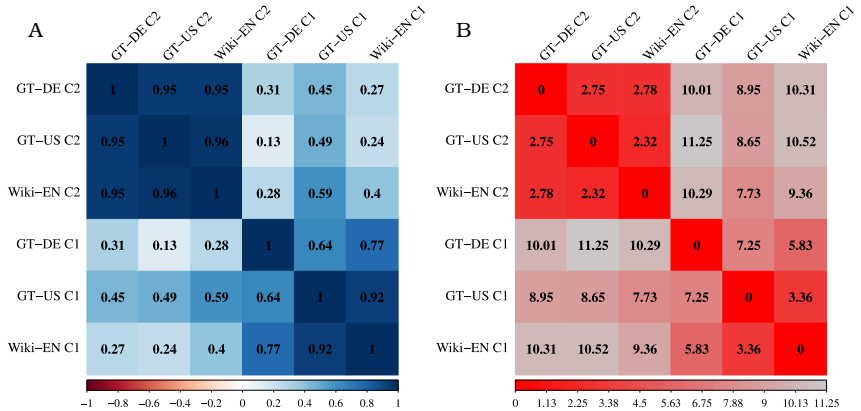

Figure 3.6: GT-US, GT-DE and Wikipedia's C1 and C2 centroids (A) correlation and (B) distance matrix. All correlations are statistically significant ($p < 0.05$). The distance matrix was computed using the Euclidean distance on z-score scaled data.

Conversely, overlapping C2 search trends are on average highly correlated (0.85 ± 0.2) and C2 centroids are very highly correlated ($R = 0.95$, $p < 0.05$).

Generally, C1 search trends include symptom-related terms, as well as terms that hint at a flu infection such as *Flu how long*. We considered the possibility that C1 search trends are measuring flu infections. We therefore collected data on the laboratory-confirmed cases as a possible explanatory variable.

Flu activity differed in both countries.

The pH1N1 epidemic curves for the US (A) and Germany (B) (Figure 3.7) display two distinct peaks, the first corresponding to the spring-summer wave and the second to the fall-winter wave. The fall-winter wave was more prominent than the spring-summer wave in both countries, however, the US spring-summer peak was considerably more pronounced than Germany's summer peak. Moreover, the pH1N1 summer wave onset started later in Germany and its fall-winter wave peaked one month ahead of the US's. This dissimilarity was not unexpected, as pH1N1 emerged in North America. In fact the two epidemic curves have a very low correlation ($R = 0.24$, $p < 0.05$). We hereby refer pH1N1 epidemic curves as *flu activity*.

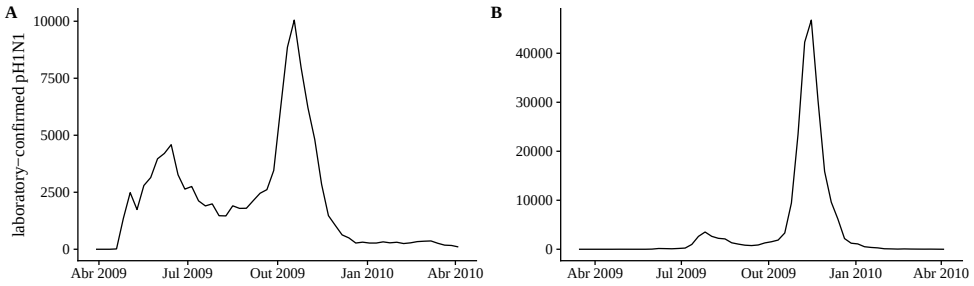

Figure 3.7: Weekly count of laboratory-confirmed pH1N1 cases in (A) United States, (B) Germany.

C1 country-specificity is explained by different flu activity.

Considering the difference between flu activity in both countries, we put the hypothesis forward that C1 series country-specificity can be explained by this difference. We tested this hypothesis by obtaining the correlations between each country's search trends and flu activity. We found that flu activity is systematically more associated with C1 search trends than C2 search trends (Figure 3.8). Moreover, C1 centroids are likewise more correlated with flu activity than C2 centroid, with a noteworthy very high correlation with GT-DE C1 centroid Table 3.2.

Table 3.2: Pearson's correlation between flu activity and cluster centroids.
* denotes *p-value < 0.05.* ** denotes *p-value < 0.001.*

R(pH1N1)	C1 Centroid	C2 Centroid
GT-US	0.73**	0.19*
GT-DE	0.91**	0.18*
Wikipedia	0.69**	0.36*

To rule out the possibility of spurious correlation we tested the presence of causal links between cluster centroids and flu activity through Granger's causality test (page 9). In Germany,

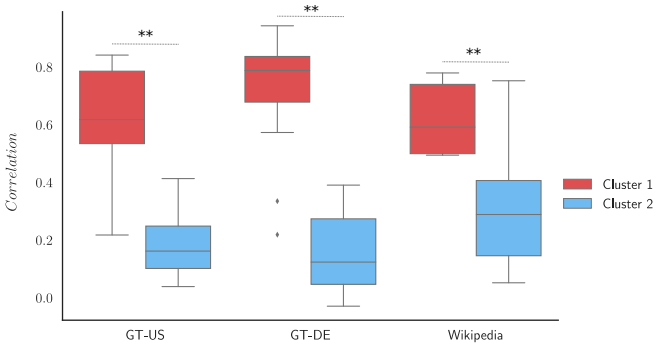

Figure 3.8: Boxplot of Pearson's correlation coefficients computed between pH1N1 cases and series contained in Cluster 1 and Cluster 2 of GT-US, GT-DE and Wikipedia regarding the whole pandemic period. ** denotes t-test rejection of H_0 of equal means, $p < 0.001$

flu activity *G-caused* GT-DE C1 centroid ($G = 9.2$, $p < 0.001$). In the United States flu activity *G-caused* GT-US C1 centroid ($G = 4.8$, $p < 0.05$) and Wiki-EN C1 centroid ($G = 9.9$, $p < 0.001$). C2 centroids were not G-caused by each respective country's flu activity (all $p > 0.05$).

Of the two uncovered patterns, just one, C1, is well explained by flu activity. The alternative pattern, C2, is seemingly unconnected to flu activity. C2 search trends are largely characterized by a prominent initial peak that quickly subsided. Even though the terms collected for C2 search trends are related to a flu pandemic, these search trends have low correlations with flu activity. We know that in the 2013 flu outbreak in the US, a high media activity resulted in abnormal search trend peaks that confounded GFT's model (Copeland *et al.*, 2013). Considering this information, we will test whether C2 search trends are associated with media activity. We collected data on media activity (measured as flu-related news counts) during the pandemic period for both Germany and the United States as a possible explanatory variables for the unexplained C2 search trends.

Media activity trended similarly in both countries.

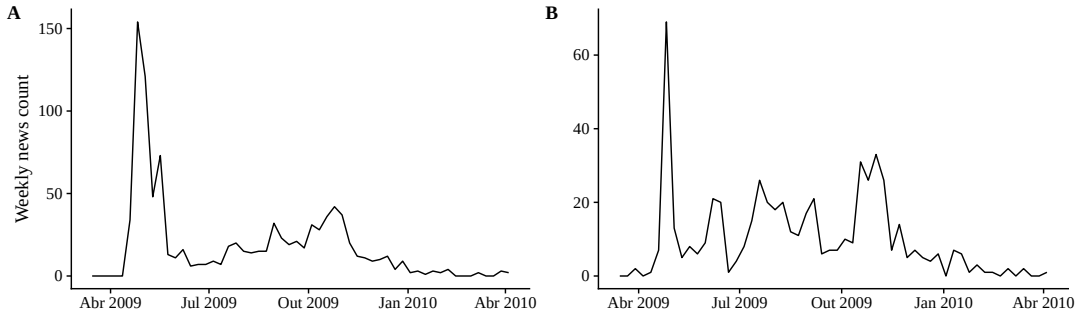

Figure 3.9: Media activity in the United States (A) and Germany (B) considering the period from March 2009 to April 2010

Media activity quickly increased, peaked, and declined during early pandemic stages, between April-May 2009, in both countries (Figure 3.9). A small surge of news is observed in early June, matching WHO's pandemic declaration. Some media activity remained afterwards, but not as intensely as initially. By the end of 2009 media activity had virtually ceased. While both country's media activity share a similar profile, judging from the collected data Germany's initial media activity was less intensive than the US's, whereas media reporting during later

stages remained similar in both countries. This can be attributable to the different collection of news sources in the two countries (Section 2.1.4). Even if the different data sources may have some effect on the comparability in absolute terms, the captured trend is comparable. In fact both country's media curves are moderately correlated ($R = 0.67$, $p < 0.001$).

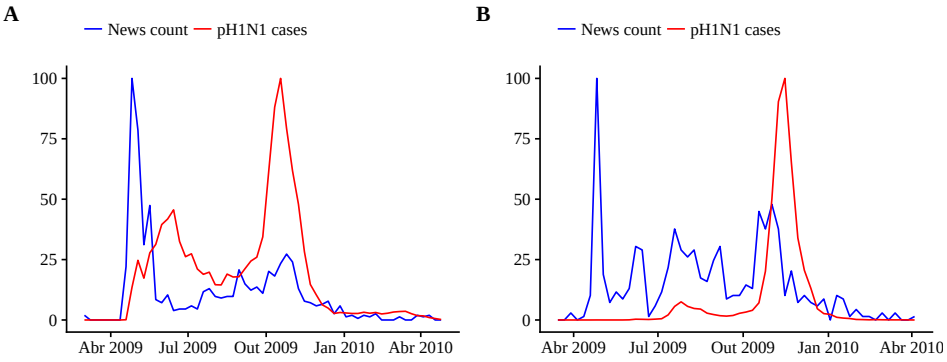

Figure 3.10: Media activity and flu activity (max-scaled) in the United States (A) and Germany (B) considering the period from March 2009 to April 2010.

Media activity peaked in Germany in April, at a time of virtually no flu activity. Moreover, analyzing the over-time dynamics between media activity and flu activity reveals a systematic asynchrony in both countries (Figure 3.10). In the US media activity reached very low levels by the time the spring-summer flu activity peaked. The fall-winter pH1N1 wave seems more in phase with US media activity, but overall both curves are unmatched. This asynchrony is further verified by the low correlations between pH1N1 and media curves in Germany ($R = 0.22$, $p < 0.05$) and the United States ($R = 0.29$, $p < 0.05$).

Even though US media activity peaks are out of phase with flu activity peaks, flu activity *Granger-caused* to some extent US media activity ($G = 2.7$, $p < 0.05$) but not the other way around ($G = 1.9$, $p > 0.05$). The same applies to Germany (*Flu activity* $\xrightarrow{G}$ *Media activity*, $G = 2.74$, $p < 0.05$). This suggests that despite the overall obvious asynchrony, media activity curves may contain some information related to flu activity.

C2 series are highly associated with media activity, but not as much with flu activity.

Considering that 1) both country's media activity is moderately similar; 2) both country's C2 centroids are highly similar; 3) flu activity is asynchronous with media activity and 4) flu activity explains C1 but not C2 search trends, we then assume that C2 search trends are possibly explained by media activity.

In order to test this assumption we computed the correlations between search trends and media activity. In addition we also computed the correlation between cluster centroids and media activity. We found that media activity is highly correlated with virtually all C2 series (Figure 3.11). C2 centroids are likewise highly correlated with media activity (Table 3.3).

Media activity also holds high correlations with some C1 search trends (Figure 3.11) and moderate correlations with C1 centroids (Table 3.3). This is attributable to C1 search trends that display a dynamic behavior in-between C1 and C2, characterized by both early and later peaks (*eg. Flu treatment* in GT-US). This is evident upon observation of the centroids (Figure 3.4). The C1 search trends that are more correlated with media activity have a low average cluster

membership ($C_m = 0.68+0.2$). Apart from these exceptions, C1 trends are overall low correlated with media activity.

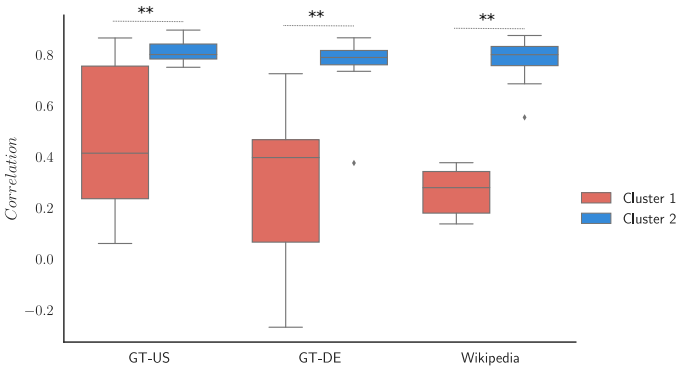

Figure 3.11: Boxplot of Pearson's correlation coefficients computed between media attention and series contained in Cluster 1 and Cluster 2 of GT-US, GT-DE and Wikipedia. ** denotes t-test $p < 0.001$

Table 3.3: Pearson's correlation between media activity and cluster centroids.
* denotes *p-value < 0.05.* ** denotes *p-value < 0.001.*

R(Media)	C1 (Centroid)	C2 (Centroid)
GT-US	0.52**	0.83**
GT-DE	0.36*	0.81**
Wikipedia	0.30*	0.85**

Moreover, Granger causality rendered concordant results, as media attention *Granger-caused* GT-US C2 centroid ($G = 11$, $p < 0.001$), GT-DE C2 centroid ($G = 9.6$, $p < 0.05$) and Wikipedia-EN C2 centroid ($G = 2.6$, $p < 0.05$), but did not *Granger-cause* any C1 centroid.

C1 is strongly linearly related to flu activity. C2 is strongly linearly related to media activity

In addition to correlations we used linear regression, with media and flu activity as explanatory variable and the search patterns as response variables. We found that C1 centroids display strong and statistically significant linear relationships with flu activity. There is also evidence of a weaker, yet statistically significant linear relationship between GT-US and Wiki-EN C1 centroids with media activity (Figure 3.12). On the other hand, C2 centroids display strong and statistically significant linear relationships with media activity, but not with flu activity.

C1 early peaks are more associated with media activity.

The distinction between C1-pH1N1 and C2-Media is not mutually exclusive. Fuzzy clustering revealed search trends that portray both C1 and C2 patterns. Since media activity was predominant on early pandemic stages, whereas flu activity was yet low, we assume some of the observed high correlations between C1 search trends with media activity are mostly due to the early media activity and not later media activity. To test this assumption we defined two periods roughly based on the two pH1N1 epidemic waves, Period 1 from April to July, 2009 (inclusive) and Period 2 from August to December, 2009 (inclusive). Period 1 is thereby characterized by

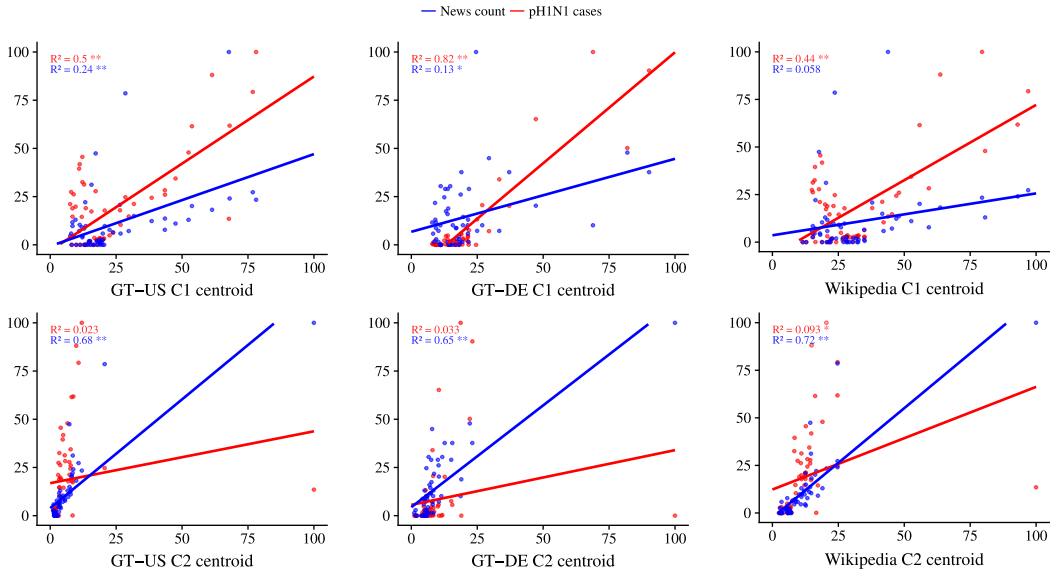

Figure 3.12: Scatterplot and regression analysis of each dataset cluster centroids against media attention curve (blue) and pH1N1 epidemic curve (red). The top-right quarter outlier in C2 represents the media activity peak; the bottom-right quarter outlier in C2 represents flu activity peak. ** denotes fit t-test $p < 0.001$, * denotes $p < 0.05$.

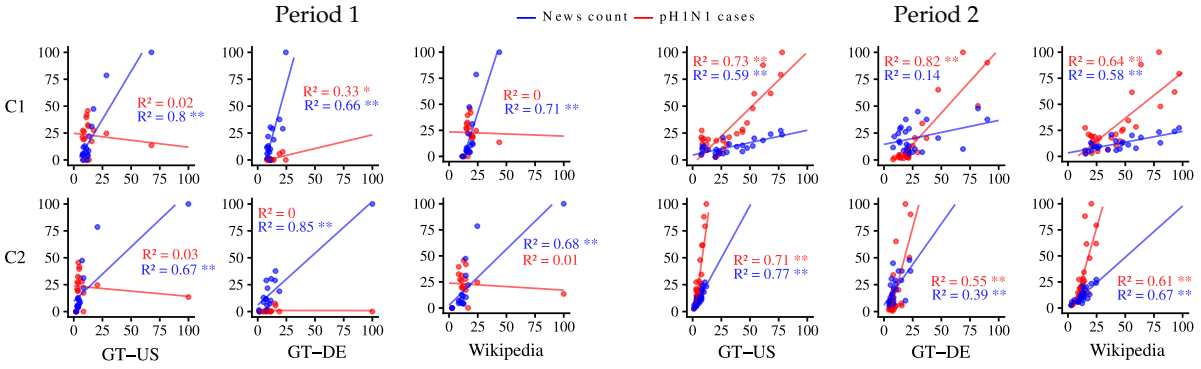

Figure 3.13: Scatterplot and segmented regression analysis of each dataset cluster centroids with media attention curve (blue) and pH1N1 epidemic curve (red). Wiki-EN centroids are max-scaled. Period 1: April-July 2009. Period 2: August-December 2009. ** denotes regression t-test $p < 0.001$, * denotes $p < 0.05$.

lower flu activity and higher media activity. On the contrary, Period 2 is characterized by higher flu activity but lower media activity.

C1 centroids display a strong linear relationship with flu activity in Period 2, but not in Period 1, except GT-DE albeit with a weak effect. On the other hand, C1 centroids have a strong linear relation with media activity in Period 1. GT-US and Wiki-EN C1 centroids also have a weaker, yet statistically significant linear relationship with media activity in Period 2 (Figure 3.12) Conversely, C2 centroids display a strong linear relationship with media activity in both Period 1 and Period 2. A weaker, yet statistically significant linear relationship is also observed between C2 centroids and flu activity in Period 2 (Figure 3.12).

We further explored the previous observation of non-matching C1-Flu activity during the spring-summer wave by observing GT-US and GT-DE C1 centroids curves together with media and flu activity in Figure 3.14. It is evident that GT-US C1 centroid early peak does not match spring-summer flu activity peak (*), but GT-US C1 matched the media activity peak instead (*p1*). GT-US C1 search trends did not respond to the pH1N1 summer peak (*). The fall-winter pH1N1 peak (*p2*) is matched by GT-US C1 later peak. (Figure 3.14).

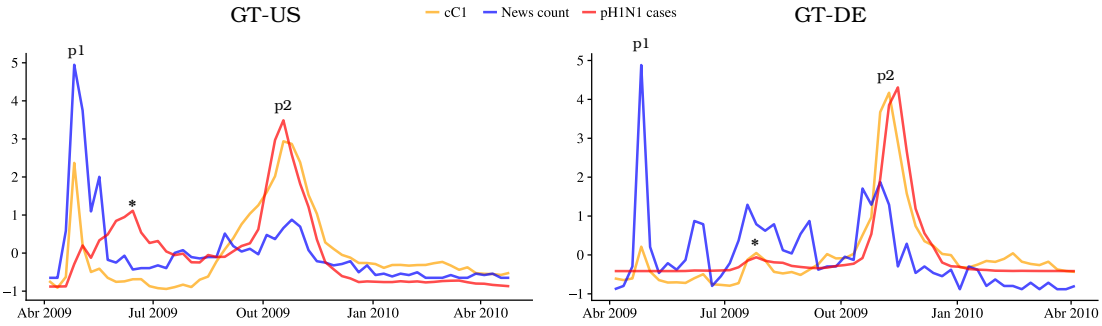

Figure 3.14: Over time dynamics (April 2009 - April 2010) between C1 centroids (cC1) of GT-US/ Wikipedia-EN (A) and GT-DE (B) with pH1N1 cases and media attention series. All series are z-score normalized. *p1* denotes the early C1 series peak; *p2* the later C1 series peak; * denotes each respective country summer-wave pH1N1 peak.

In Germany C1 search trends matched the pH1N1 summer peak (*). Despite having virtually no flu activity by late April, GT-US C1 slightly peaked in parallel with media activity peak (*p1*). This early peak (p1) is similar in magnitude to the to the matched C1/flu activity peak (*).

Anxiety levels paralleled media activity. Perceived pH1N1 risk paralleled flu activity.

To know the the public's anxiety and risk perception varied along the pandemic we estimated anxiety and risk perception levels from several surveys as described in page 7. Risk perception was initially low. Risk perception started to increase by July 2009 and peaked on October 2009. Anxiety, on the other hand, was initially high but quickly subsided, reaching low levels by July 2009 (Figure 3.15).

Anxiety and Risk perception appear to be inversely related. In fact, Anxiety and Risk perception are moderately and negatively correlated, albeit statistically insignificant ($R = -0.45$, $p\text{-}value$=0.19) due to the small sample. Based on previous findings, anxiety is possibly associated with media (Jones & Salathé, 2009) and risk perception is possibly associated with flu activity (Gidengil *et al.*, 2012). As we obtained monthly estimates for risk perception and anxiety, we aggregated the remaining datasets from weeks to months in order make them comparable. We

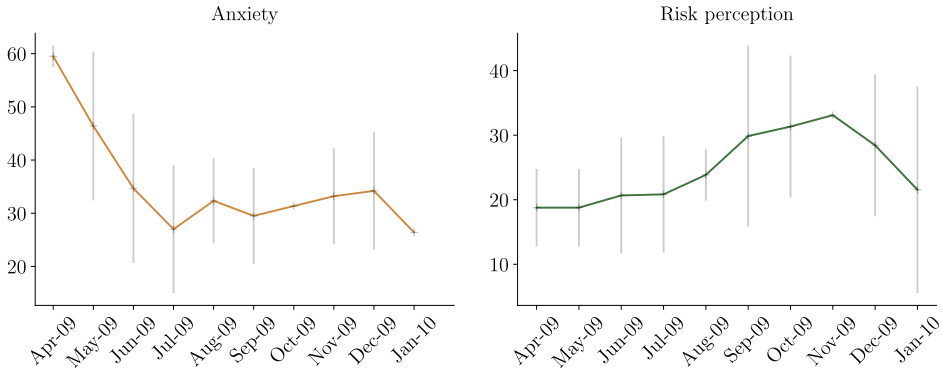

Figure 3.15: Anxiety and Risk perception estimate over time (April 2009 - January 2010), obtained by computing the mean of extracted values from collected surveys page 7). All months are covered by at least one survey, except October-09 in Anxiety. Vertical lines denote standard deviation.

verified both associations by obtaining risk perception and anxiety correlations with the US/DE averaged media activity and US/DE averaged flu. We averaged US/DE media and US/DE flu activity, as anxiety and risk perception estimates were obtained from several different countries. Risk perception is highly and positively correlated US/DE average flu activity ($R = 0.7$, $p < 0.05$), but not with the averaged media activity ($R = -0.04$, $p > 0.05$). Conversely, Anxiety is highly and positively correlated with US/DE averaged media activity ($R = 0.74$, $p < 0.05$) and not with US/DE average pH1N1 activity ($R = -0.21$, $p > 0.05$).

C1 proxied Risk perception. C2 proxied Anxiety.

To infer whether online search trends proxied the estimated anxiety and risk perception real-life behaviors we computed each C1 and C2 search trend correlation with Anxiety and Risk perception (Figure 3.16). As we only have a sample of $n = 10$ we did not consider a correlation t-test in this analysis, but individual correlations and statistical significance can be accessed in pages 29 to 31. Additionally we used a linear regression with Anxiety and Risk Perception as explanatory variables and C1/C2 centroids as response variables and statistical significance is provided in this test (Figure 3.17).

Our results show that C1 search trends are not correlated with Anxiety, but are highly correlated with Risk Perception (Figure 3.16). C1 centroids are not linearly related to Anxiety but C1 centroids have strong and statistical significant linear relationship with Risk perception (Figure 3.17). C2 search trends are highly correlated with Anxiety, but not with Risk Perception (Figure 3.16). C2 centroids have a strong and statistically significant linear relationship with Anxiety, but not Risk perception(Figure 3.17). Therefore, C1 search trends proxied Risk Perception and C2 search trends proxied Anxiety.

Results overview

We provide an aggregated overview the results, including each individual series across the three datasets in Figure 3.18 (GT-US), Figure 3.19 (GT-DE) and Figure 3.20 (Wiki-EN).

In GT-US the overall separation between C1-Flu and C2-Media is evident. A set of series indicated by lower cluster membership levels are equally explained in terms of correlations and distance by both media and flu activity. Nevertheless these series are more *G-caused* by media activity than flu activity. Regardless, the Granger causality test corroborates that Media activity

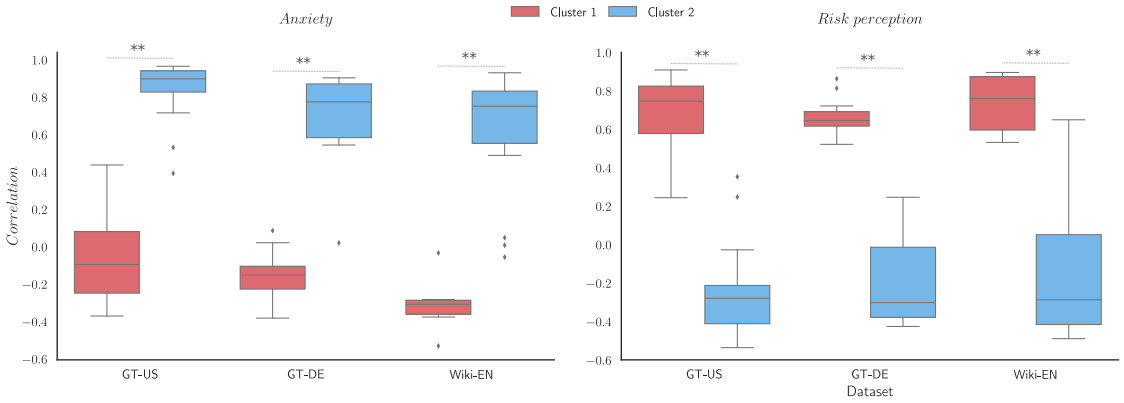

Figure 3.16: Boxplot of correlations between C1/C2 series and Anxiety/Risk perception. For comparability's sake, considering $n = 10$, just correlation coefficients but not their statistical significance are taken into account. Statistical significance for each search trend is shown in pages 29 to 31. ** denotes t-test *p-value* < 0.001

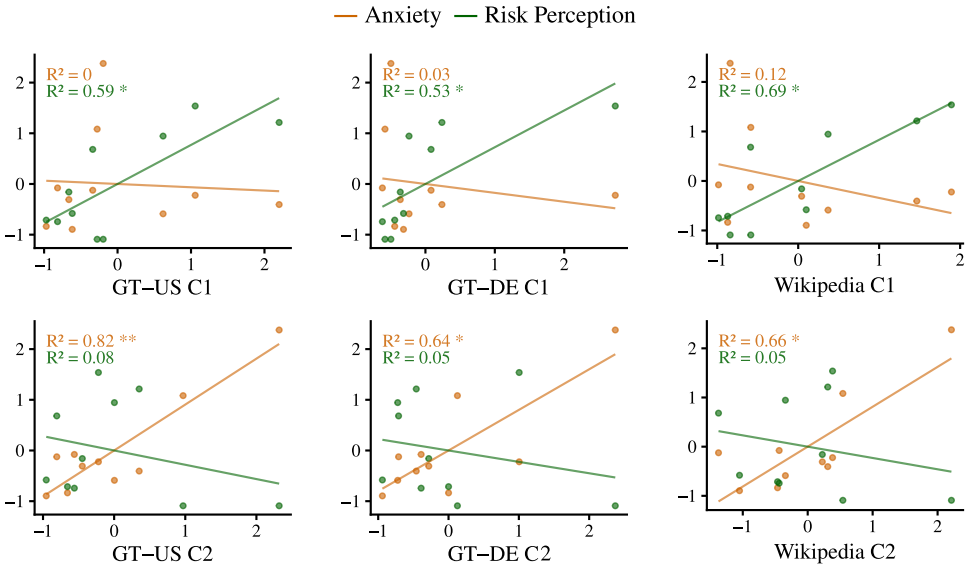

Figure 3.17: Scatterplot and regression line of cluster centroids with anxiety and Risk perception levels. Values are z-score normalized. Each monthly dataset cluster centroid is shown in page 60. ** denotes regression t-test $p < 0.001$, * denotes $p < 0.05$.

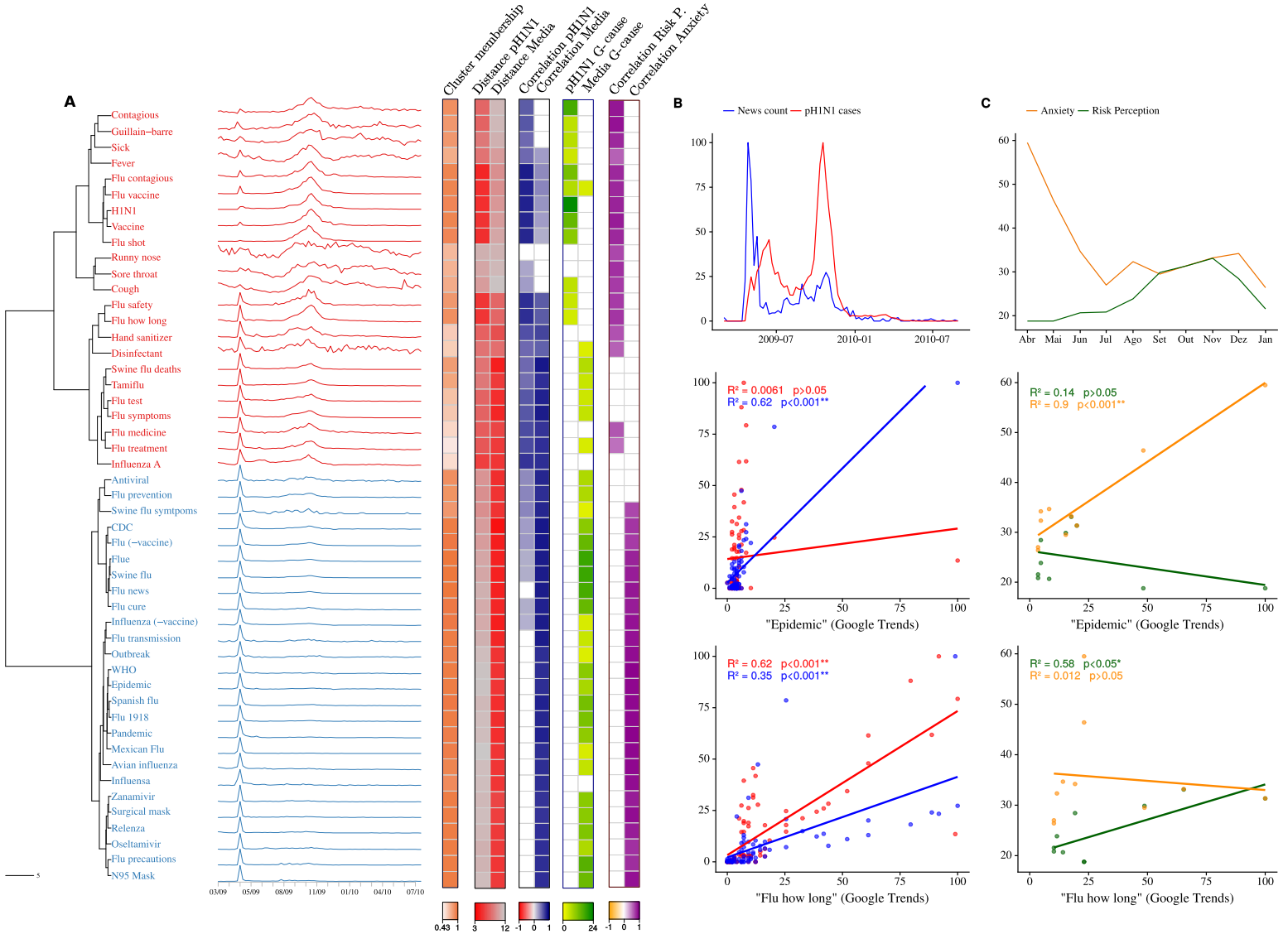

Figure 3.18: GT-US results overview. [A] is the dendrogram with Cluster 1 (Red) and Cluster 2 (Blue); Search trends are shown in the middle column. The right columns is each search trend's respective analysis. [B] and [C] show Media/Flu effect and Anxiety/Risk perception effect on selected terms from each cluster. Blank squares in [A] indicate no statistically significant result ($\alpha = 0.05$, except risk perception and anxiety correlations ($\alpha = 0.10$).

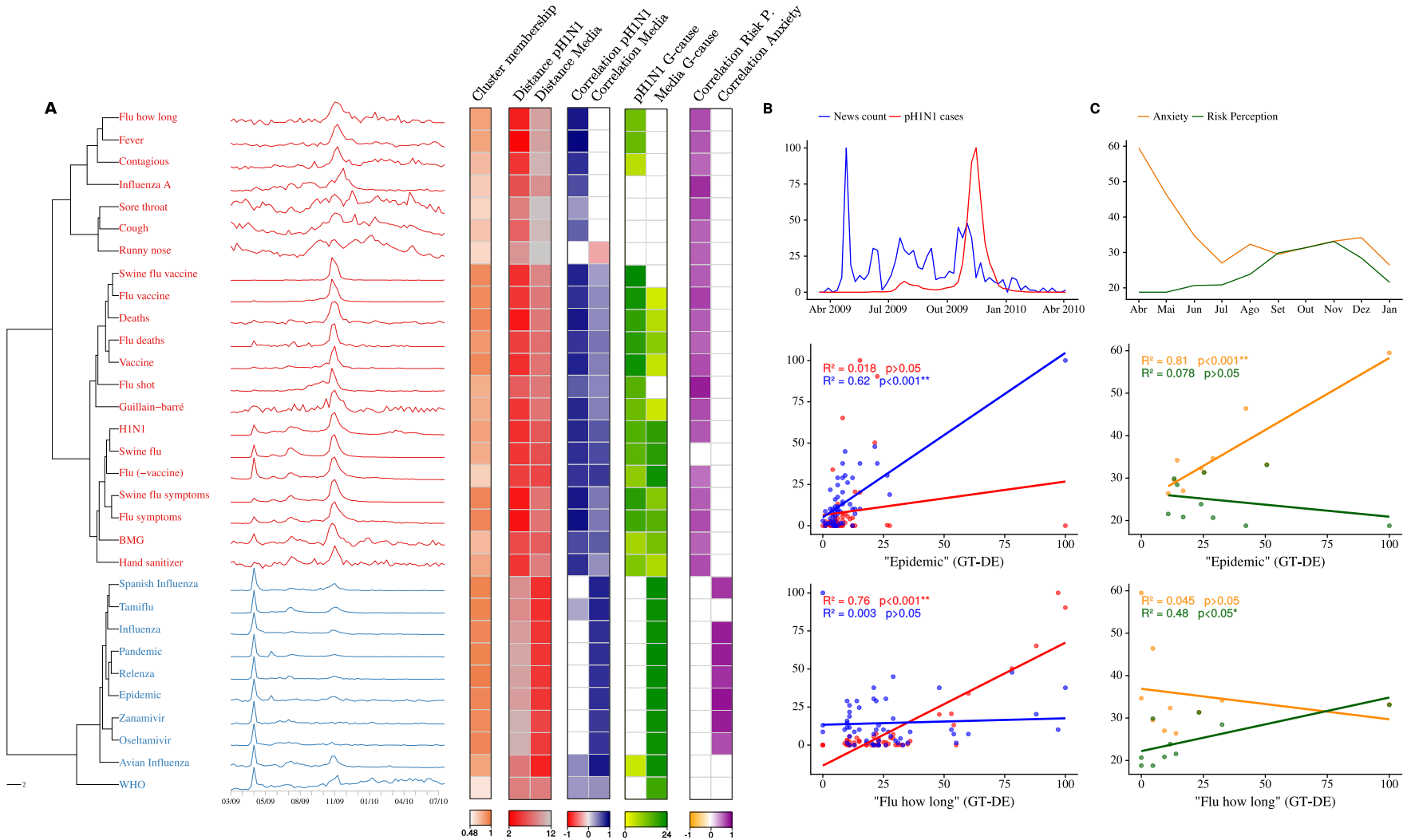

Figure 3.19: GT-DE results overview. [A] is the dendrogram with Cluster 1 (Red) and Cluster 2 (Blue); Search trends are shown in the middle column. The right columns is each search trend's respective analysis. [B] and [C] show Media/Flu effect and Anxiety/Risk perception effect on selected terms from each cluster. Blank squares in [A] indicate no statistically significant result ($\alpha = 0.05$, except risk perception and anxiety correlations ($\alpha = 0.10$).

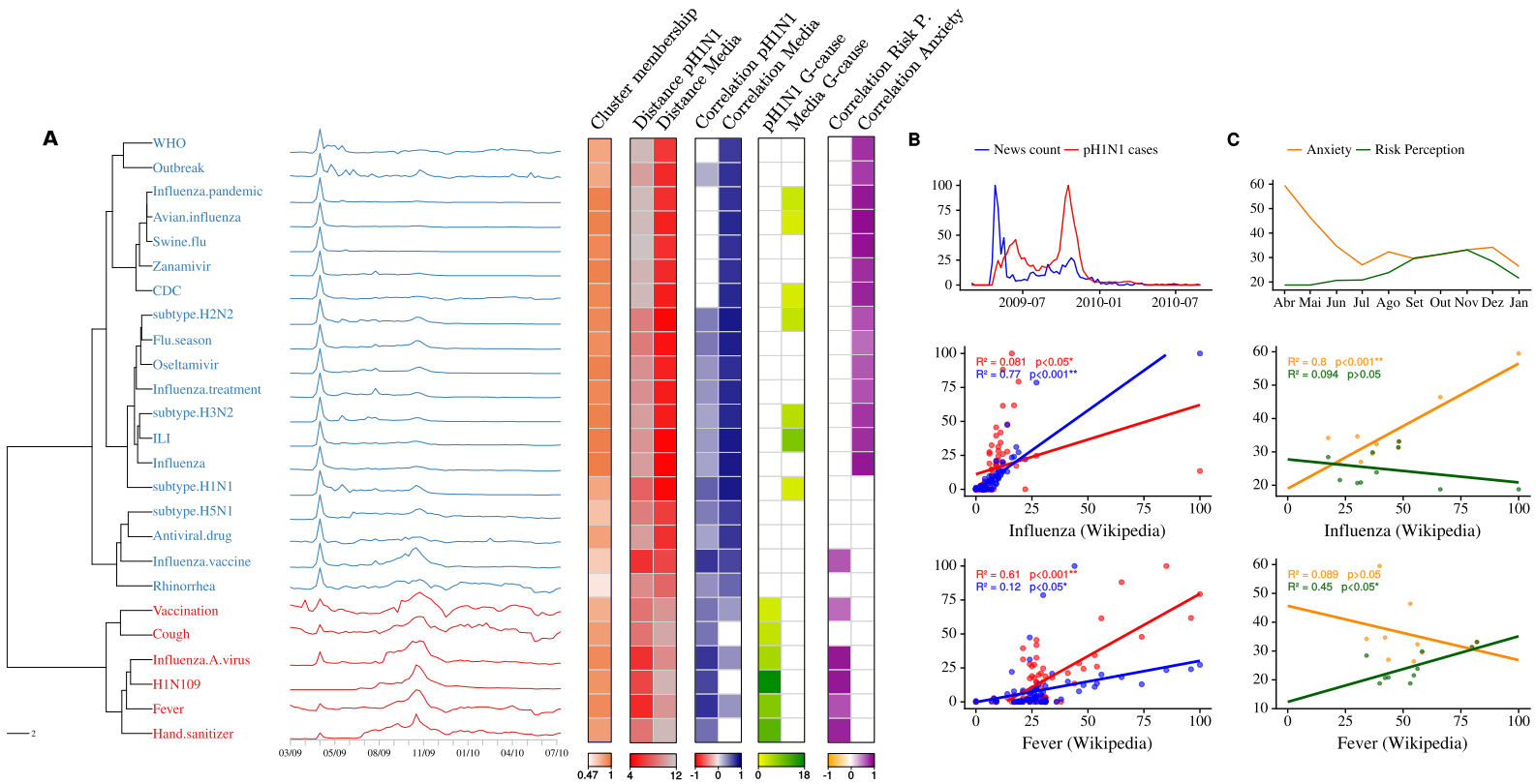

Figure 3.20: Wiki-EN results overview. [A] is the dendrogram with Cluster 1 (Red) and Cluster 2 (Blue); Search trends are shown in the middle column. The right columns is each search trend's respective analysis. [B] and [C] show Media/Flu effect and Anxiety/Risk perception effect on selected terms from each cluster. Blank squares in [A] indicate no statistically significant result (($\alpha = 0.05$, except risk perception and anxiety correlations ($\alpha = 0.10$).

Granger-causes virtually all C2 search trends. On the other hand, Flu activity *Granger-causes* most C1 search trends. C1 search trends are more correlated with Risk perception; C2 search trends are more correlated with Anxiety. In B) and C) we provide two selected series, one of each cluster, to display the explanatory variables effect in a more individual term setting. The *Epidemic* search trend is very strongly and linearly related with media activity, but not with flu activity. Additionally, *Epidemic* search trend is strongly related to Anxiety but not Risk perception. On the other hand, the *Flu how long* search trend conveys a strong linear relationship with flu activity, and a weaker yet relevant linear relationship with media activity. This search trend is associated with Risk perception, but not Anxiety. The remaining scatterplots and regression results for each GT-US term are supplemented in pages 49 to 52.

GT-DE results display an overall similar scenario to GT-US results, with a clear distinction between C1-flu and C2-media and few C1 series being also correlated with media activity. Some differences are also evident. Contrasting with GT-US C1 series, the early peak in GT-DE C1 series is not as prominent. This is made evident by comparing the provided examples in GT-US with GT-DE (Figure 3.18C, Figure 3.19C), as the same search trend for *Flu how long* display a different media effect in both countries. In GT-DE this term has linear relationship with with media activity, whereas in GT-US it displays a moderate and statistically significant relationship with media activity. Nonetheless, anxiety and risk perception results are the same in both countries. The remaining scatterplots and regression results for each GT-DE term are supplemented in pages 53 to 56.

Wiki-EN results are overall concordant with the remaining datasets. Although it is the smallest dataset, the distinction between C1-flu-Risk perception and C2-media-anxiety remains evident, however unlike in GT-US and GT-DE, Granger causality did not detect statistically significant media activity causal links to some C2 series, despite the high correlations. Considering Wiki-EN is not exclusively used by US inhabitants, then this observation is possibly due to noise introduced by non-US individuals. The provided examples for Wiki-EN portray a similar situation, with a C1 search trend *Fever* being strongly and linearly related to flu activity and risk perception, and on the other hand, a C2 search trend *Influenza* being strongly and linearly related to media activity and anxiety. The remaining scatterplots and regression results for each Wiki-EN term are supplemented in pages 57 to 60.

Chapter 4

Discussion

In this work, we have presented data showing that 1) flu-related search trends differed between the pandemic and the seasonal peripandemic periods; 2) two prominent peaks with varying magnitudes were identified in pandemic search trends 3) pandemic search trends are clustered into two major groups; 4) one group strongly correlates with flu activity in each country and risk perception; 5) the other group strongly correlates with media activity and with self-reported anxiety; and 6) these patterns are visible in at least two countries and in two different online platforms.

This suggests that it should be possible to identify words and search terms that, despite being *a priori* similar and related, are in fact independent and might inform on different behaviours.

We think that these results are important for several reasons. That different flu-related search terms display significantly different patterns can be used in the public health setting, helping to monitor and manage both disease and concern. Terms such as *H1N1* or *Contagious* that strongly correlate with the actual number of influenza reported cases, but that seem to be less sensitive to media hype, could be used in disease tracking and surveillance (after further validation). The other search terms (*eg. Spanish flu*, *Pandemic*) are strongly correlated with anxiety and media activity, but that do not relate with actual number of influenza cases, thus can be used to better monitor population anxiety levels and manage risk communication. Moreover, this work provides a proof of principle analysis of connecting online with real-life behaviour in a setting previously believed to be unique and intractable.

In fact, that the public's online response to the pandemic disturbed the seasonality of search trends (considered to be reliable measures of flu cases) was previously known (Cook *et al.*, 2011) and the main reason why the 2009-2010 flu season is not usually included in flu surveillance models based on online data (Hickmann *et al.*, 2014; McIver & Brownstein, 2014; Won *et al.*, 2017).

In addition, the pandemic led to a sudden interest in specific search trends that almost disappeared afterwards, for instance the *Tamiflu* and *Relenza* search trends in both Germany and US. Likewise, media activity was at its highest during late April 2009, but thereafter quickly subsided to low values, slightly increasing again by the fall-winter wave. This pattern was very similar in both Germany and the US, and there is evidence that it would be found in other countries (Smith *et al.*, 2013). As we uncovered a set of search trends that followed this exact same pattern (also similar between the two countries), with a prominent early peak and low levels of interest thereafter, we suggest that these can be used as "media-associated search trends" terms, unrelated with the actual number of cases. And this is unlikely to be specific

to search terms as a similar pattern was observed in Twitter data (Chew & Eysenbach, 2010; Signorini *et al.*, 2011).

As flu activity differed in both countries, so did a specific set of search trends. These search trends (*eg. Flu symptoms* in the US) adequately measured flu activity during the fall-winter but not during the spring-summer wave, particularly in the US. In fact, they were more associated with media activity during the spring-summer wave than with flu activity. In addition, they peaked in parallel with the media activity peak during late April 2009. Taking into account that this was a period of low flu incidence, it is then unlikely that these search peaks resulted from the activity of pandemic flu infected individuals. And while our analysis did not support a strong quantitative media effect on flu-measuring search trends during the fall-winter wave, it should not be disregarded, as for instance terms such as *Deaths* or *Swine flu deaths* peaked in the fall-winter wave. This suggests that we possibly lost information by just considering the news counts and not news content. Media activity in terms of news counts is a strong explanatory variable of some search trends, but it appears that even with low media activity, news content effect is significant enough to cause online interest. Consistently, we found that media activity series contained information regarding flu activity, despite both series having low correlation.

Our data also shows that the public's anxiety was associated with media activity and that both anxiety levels and media activity were at their highest level in late April 2009, when the infection-hinting search trends also peaked. Media's unfolding of the pandemic crisis caused manifestations of anxiety (Jones & Salathé, 2009) and this in turn motivated infection-hinting searches unrelated with actual flu infections but very likely related with the public's anxiety instead. This can justify the observed underperformance of well established flu-measuring search trends during the 2009 pandemic. The media effect on these search trends likely overshadowed the true signal of infection.

Why presumably non-infected individuals searched for terms that suggested a flu infection remains to be explained. It could be due to 1) one's misjudgment of one's own symptoms as a pandemic flu infection, 2) due to the occurrence of psychogenic illness, where individuals experience symptoms that have no discernible physical cause (Bass *et al.*, 2012), 3) because people are trying to anticipate a possible infection and want to be prepared, 4) because they are curious and want more information, or 5) because they are looking for information for someone else. In fact, a similar pattern to what we observed in these search trends was also observed in a real-world context, where emergency departments experienced substantial increases in patient volumes at a time of high media activity but low flu activity (Codish *et al.*, 2014; Keramarou *et al.*, 2011; McDonnell *et al.*, 2012).

However, we were also able to identify search terms that are uniquely associated with the number of cases that appear to be less sensitive to the media, and that do not correlate with anxiety levels (eg. *H1N1* in the US). These could, in principle, be used in surveillance systems, after further curation and validation. It is important to point out that search trends that highly correlate with flu activity cannot be assumed to have resulted merely from the collective activity of flu infected individuals. For example, vaccine-related search trends are strongly correlated with flu-measuring search trends, but vaccine information-seeking is pointless from the infected individuals point of view and it is unlikely that these vaccine-related search trends resulted from online activity of infected individuals. Moreover, the widespread flu activity during the fall winter-wave increased the likelihood of flu infection, and in in turn, the public's perception of this likelihood also increased, either due to 1) media coverage of increasing flu cases or 2) due to

knowing someone infected. In fact, a survey conducted during the fall-winter wave found that on average 25% of the respondents knew someone infected with the pandemic flu, up from 6% in the spring-summer wave (Caravan, 2009). Thus, we found that the set of search trends associated with flu activity not only measured this flu activity but also proxied the public's perception of the flu activity.

Taken together our results support the hypothesis that it should possible to identify good proxies of offline behaviour in online data and that these tools can become very relevant in modern epidemiology and public health. Moreover, and despite focusing on the flu pandemic, the principles of our analysis should be applicable to other health crisis settings and possibly even to non-health related situations.

4.1 Limitations

As we are dealing with a very complex setting, involving the individual actions of large numbers of people, that cannot be validated, there are several confounding variables that must be taken into account. In addition to the ones mentioned throughout the document, we would like to list some other limitations of this study.

Regarding the number of flu cases, as over-reporting is more prone to occur during public health crisis, there could have been reporting asymmetries between both epidemic waves (White & Pagano, 2010). Regardless, the laboratory-confirmed cases should at least reflect how pH1N1 infections trended over time and it is very unlikely that the trend should be reversed.

Regarding the search-terms data, there are several limitations that there mentioned before: Google does not provide an absolute count of search queries, offering instead a normalized trend. As we studied normalized search query temporal patterns, no comparisons in terms of magnitude can be made between the collected search trends. Normalization would pull less searched trends down, so we opted for search trends instead of magnitude (Stephens-Davidowitz & Varian, 2014). Wikipedia does not offer geo-located data and the only information we could gather is on the language used to do the search. It is unlikely that Wikipedia's time series are independent to Google's, since Google searches represent a main source of volume into Wikipedia (`https://stats.wikimedia.org/wikimedia/squids/SquidReportOrigins.htm`). Moreover, we could not get access to the German searches so this analysis is limited to the US.

Our approach also depended on a collection of search trends that should only apply in to context of a flu pandemic. However, the collection of circumstantial terms related to other health crisis settings should be as straightforward as was in this context.

Regarding the media, the US and German datasets are difficult to compare directly, as they include very disparate numbers of news sources. We only counted the newspaper articles and television broadcasts about pH1N1 in Germany which had been collected in context of another study (Reintjes *et al.*, 2016). Additionally, different sources were used in Germany and US analysis. However, it is unlikely that adding more news sources to the German dataset would alter ous analysis, as comparison of different media sources tends to find little difference between them (Smith *et al.*, 2013). Another limitation was that our analysis was blind to news content.

Regarding the survey dataset, by aggregating data from several different surveys we are possibly including biases that are intrinsic to different methodologies (Blumberg & Luke, 2007; Duffy *et al.*, 2005). Moreover, different questions were asked to the respondents in the Anxiety category. The lack of standardized survey formats led to different surveys asking respondents

about their 'concern', 'worry' or 'anxiety', which are frequently used interchangeably. However, different phrasing around the same concept may lead to different answers (Consedine *et al.*, 2004). We also collected surveys from different countries, which should add to the observed variability. Yet, Tooher *et al.* (2013) reviewed several pandemic-related surveys and found very low inter-country variability in the public's response.

Finally, and regarding our analysis, it is very difficult to establish causality and it is important to point out that we are only identifying correlations and relationships between variables, albeit strong and significant.

4.2 Conclusion

In this work we provided a proof of principle that search trends online-based surveillance models limitation can be surpassed by pinpointing what motives are likely at stake. In addition, this limitation also provided an excellent opportunity to understand real-life human behavior through online data. Our findings further support the usefulness of online data to understand real-life behaviors, but also make evident the difficulty in overcoming inherent limitations associated with this type of data.

4.3 Future work

To be able to generalize these results, the analysis should be extended to include more countries, as long as it is possible to collect their respective media and flu activity datasets.

We can make specific predictions that should be testable. For instance, by monitoring the performance (in terms of measuring flu activity) of search trends that are less sensitive to media effect against trends more sensitive to media effect we can validate this approach.

It would also be very interesting to include user-generated content, such as Twitter posts. This is more complex than retrieving search trends, yet this complexity pays off as it offers the potential of distinguishing what motivated users to share something about the flu (Lamb *et al.*, 2013): an individual may tweet about being infected, or just share news about pandemic developments, for instance.

Finally, it would be very interesting to test whether the principle of online behaviors proxying real-life behaviors applies to other public health crisis, such as the 2014 Ebola outbreak, and to what extent media and epidemic curves modulated such behaviors. In the long run, it might be possible to test whether this principle can be extended to a broader context, for instance the public's reaction to an economic crisis.

References

Agüero, F., Adell, M.N., A, A.P.G., Medina, J.L. & X, X.G.C. (2011). Attitudes and Preventive Behaviours Adopted During the (H1N1) 2009 Influenza Virus Epidemic in Spain. 73–80. 65

Anzur, T. (2000). How to talk to the media: televised coverage of public health issues in a disaster. *Prehospital and disaster medicine*, **15**, 70–72. 3

Bass, E., Kaplan-Liss, E., Dorf, D. & Broderick, J.E. (2012). A challenging empirical question: what are the effects of media on psychogenic illness during a community crisis? *Journal of community medicine & health education*, **2**. 34

Bezdek, J.C., Ehrlich, R. & Full, W. (1984). Fcm: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, **10**, 191–203. 13

Blumberg, S.J. & Luke, J.V. (2007). Coverage bias in traditional telephone surveys of low-income and young adults. *Public Opinion Quarterly*, **71**, 734–749. 35

Brammer, L., Blanton, L., Epperson, S., Mustaquim, D., Bishop, A., Kniss, K., Dhara, R., Nowell, M., Kamimoto, L. & Finelli, L. (2011). Surveillance for influenza during the 2009 influenza a (H1N1) pandemic-United States, April 2009-March 2010. *Clinical Infectious Diseases*, **52**, 27–35. 3

Bults, M., Beaujean, D.J., de Zwart, O., Kok, G., van Empelen, P., van Steenbergen, J.E., Richardus, J.H. & Voeten, H.A. (2011). Perceived risk, anxiety, and behavioural responses of the general public during the early phase of the Influenza A (H1N1) pandemic in the Netherlands: results of three consecutive online surveys. *BMC public health*, **11**, 2. 64, 65

Caravan (2009). H1n1 flu preparedness (http://mabas.org/lists/announcements/attachments/56/flupreppoll.pdf). 35

Chan, T.C., Fu, Y.c., Wang, D.W. & Chuang, J.H. (2014). Determinants of receiving the pandemic (h1n1) 2009 vaccine and intention to receive the seasonal influenza vaccine in taiwan. *PLOS ONE*, **9**, 1–9. 4

Chew, C. & Eysenbach, G. (2010). Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. *PLoS ONE*, **5**, e14118. 34

Choi, H. & Varian, H. (2012). Predicting the present with google trends. *Economic Record*, **88**, 2–9. 2

Codish, S., Novack, L., Dreiher, J., Barski, L., Jotkowitz, A., Zeller, L. & Novack, V. (2014). Impact of mass media on public behavior and physicians: an ecological study of the h1n1 influenza pandemic. *Infection Control & Hospital Epidemiology*, **35**, 709–716. 34

Consedine, N.S., Magai, C., Krivoshekova, Y.S., Ryzewicz, L. & Neugut, A.I. (2004). Fear, anxiety, worry, and breast cancer screening behavior: a critical review. *Cancer Epidemiology and Prevention Biomarkers*, **13**, 501–510. 36

Cook, S., Conrad, C., Fowlkes, A.L. & Mohebbi, M.H. (2011). Assessing Google Flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. *PLoS ONE*, **6**, 1–8. 2, 33

Copeland, P., Romano, R., Zhang, T., Hecht, G., Zigmond, D. & Stefansen, C. (2013). Google disease trends: an update. *Nature*, **457**, 1012–1014. 2, 22

Dawood, F.S., Iuliano, A.D., Reed, C., Meltzer, M.I., Shay, D.K., Cheng, P.Y., Bandaranayake, D., Breiman, R.F., Brooks, W.A., Buchy, P. *et al.* (2012). Estimated global mortality associated with the first 12 months of 2009 pandemic influenza a h1n1 virus circulation: a modelling study. *The Lancet infectious diseases*, **12**, 687–695. 3

Devaux, I., Kreidl, P., Penttinen, P., Salminen, M., Zucs, P., Ammon, A., influenza surveillance group, E. & coordinators for influenza surveillance, N. (2010). Initial surveillance of 2009 influenza A(H1N1) pandemic in the European Union and European Economic Area, April-September 2009. *Euro surveillance : bulletin europ??en sur les maladies transmissibles = European communicable disease bulletin*, **15**, 1–11. 3

Duffy, B., Smith, K., Terhanian, G. & Bremer, J. (2005). Comparing data from online and face-to-face surveys. *International Journal of Market Research*, **47**, 615. 35

Duncan, B. (2009). How the media reported the first days of the pandemic (H1N1) 2009: results of EU-wide media analysis. *Euro surveillance : bulletin europ??en sur les maladies transmissibles = European communicable disease bulletin*, **14**, 19286. 2, 3

Dunn, J.C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, **4**, 95–104. 12

Eastwood, K., Durrheim, D.N., Butler, M. & Jones, A. (2010). Responses to pandemic (H1N1) 2009, Australia. *Emerging Infectious Diseases*, **16**, 1211–1216. 64

Eysenbach, G. (2002). Infodemiology: The epidemiology of (mis) information. *The American journal of medicine*, **113**, 763–765. 1

Fenichel, E.P., Castillo-Chavez, C., Ceddia, M.G., Chowell, G., Parra, P.A.G., Hickling, G.J., Holloway, G., Horan, R., Morin, B., Perrings, C., Springborn, M., Velazquez, L. & Villalobos, C. (2011). Adaptive human behavior in epidemiological models. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 6306–11. 4

Ferrante, G., Baldissera, S., Moghadam, P.F., Carrozzi, G., Trinito, M.O. & Salmaso, S. (2011). Surveillance of perceptions, knowledge, attitudes and behaviors of the Italian adult population (18-69 years) during the 2009-2010 A/H1N1 influenza pandemic. *European Journal of Epidemiology*, **26**, 211–219. 64, 65

Flahault, A., Dias-Ferrao, V., Chaberty, P., Esteves, K., Valleron, A. & Lavanchy, D. (1998). Flunet as a tool for global monitoring of influenza on the web. *JAMA*, **280**, 1330–1332. 5, 7

Fowlkes, E.B. & Mallows, C.L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, **78**, 553–569. 12

Fox, S. (2006). *Online health search 2006*. Pew Internet & American Life Project. 1

Galili, T. (2015). dendextend: an r package for visualizing, adjusting, and comparing trees of hierarchical clustering. *Bioinformatics*. 12

Gallup (2013). Flash eurobarometer 287. 64, 65

Garten, R.J. & Davis, N.J., Cox (2009). Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. *Science (New York, N.Y.)*, **325**, 197–201. 3

Gaygisiz, Ü., Gaygisiz, E., Özkan, T. & Lajunen, T. (2012). Individual differences in behavioral reactions to h1n1 during a later stage of the epidemic. *Journal of infection and public health*, **5**, 9–21. 4

Gidengil, C.A., Parker, A.M. & Zikmund-Fisher, B.J. (2012). Trends in risk perceptions and vaccination intentions: A longitudinal study of the first year of the H1N1 pandemic. *American Journal of Public Health*, **102**, 672–679. 26, 65

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S. & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, **457**, 1012–4. 1

Goodwin, R., Haque, S., Neto, F. & Myers, L.B. (2009). Initial psychological responses to influenza a, h1n1 (" swine flu"). *BMC Infectious Diseases*, **9**, 166. 64

Granger, C.W.J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, **37**, 424–438. 9

GripeNet (2009). Os portugueses e a pandemia. (http://www.gripenet.pt/pt/resultados/estudo-pandemia-2009/). 64, 65

Harvard (2009). National survey finds six in ten americans believe serious outbreak of influenza a (h1n1) likely in fall/winter (https://www.hsph.harvard.edu/news/press-releases/files/swine$_flu_topline_7$.15.09.*pdf*). 64

Hickmann, K.S., Fairchild, G., Priedhorsky, R., Generous, N., Hyman, J.M., Deshpande, A. & Valle, S.Y.D. (2014). Forecasting the 2013 – 2014 Influenza Season using Wikipedia. 1–23. 1, 6, 33

Iksoon, E. (1996). A note on derivation of the least squares estimator. Tech. rep. 8

Jefferson, T., Foxlee, R., Mar, C.D., Dooley, L., Ferroni, E., Hewak, B., Prabhala, A., Nair, S. & Rivetti, A. (2008). Physical interventions to interrupt or reduce the spread of respiratory viruses: systematic review. *BMJ*, **336**, 77–80. 3

Jones, J.H. & Salathé, M. (2009). Early assessment of anxiety and behavioral response to novel swine-origin influenza a(H1N1). *PLoS ONE*, **4**, 2–9. 4, 26, 34, 65

Kelley, K., Clark, B., Brown, V. & Sitzia, J. (2003). Good practice in the conduct and reporting of survey research. *International Journal for Quality in Health Care*, **15**, 261–266. 4

Keramarou, M., Cottrell, S., Evans, M.R., Moore, C., Stiff, R.E., Elliott, C., Thomas, D.R., Lyons, M. & Salmon, R.L. (2011). Two waves of pandemic influenza A(H1N1) 2009 in Wales–the possible impact of media coverage on consultation rates, April-December 2009. *Euro surveillance : bulletin européen sur les maladies transmissibles = European communicable disease bulletin*, **16**, 1–7. 34

Kim, M. & Ramakrishna, R.S. (2005). New indices for cluster validity assessment. *Pattern Recogn. Lett.*, **26**, 2353–2363. 12

Klemm, C., Das, E. & Hartmann, T. (2014). Swine flu and hype: A systematic review of media dramatization of the H1N1 influenza pandemic. *Journal of Risk Research*, **9877**, 37–41. 2, 3

Krause, G. (2010). Rückblick: Epidemiologie und infektionsschutz im zeitlichen verlauf der influenzapandemie (h1n1) 2009. 5, 7

Lamb, A., Paul, M.J. & Dredze, M. (2013). Separating fact from fear: Tracking flu infections on Twitter. *Proceedings of NAACL-HLT 2013*, 789–795. 1, 36

Lazer, D., Kennedy, R., King, G. & Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science*, **343**, 1203–1205. 2

Little, T.D. (2013). *Longitudinal structural equation modeling*. Guilford Press. 7

Liu, Y., Li, Z., Xiong, H., Gao, X. & Wu, J. (2010). Understanding of internal clustering validation measures. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, 911–916, IEEE. 12

Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K., Aboyans, V., Abraham, J., Adair, T., Aggarwal, R., Ahn, S.Y. *et al.* (2012). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the global burden of disease study 2010. *The lancet*, **380**, 2095–2128. 1

Ma, R. (2005). Media, crisis, and sars: An introduction. *Asian Journal of Communication*, **15**. 3

McDonnell, W.M., Nelson, D.S. & Schunk, J.E. (2012). Should we fear "flu fear" itself? effects of h1n1 influenza fear on ed use. *The American Journal of Emergency Medicine*, **30**, 275 – 282. 34

McIver, D.J. & Brownstein, J.S. (2014). Wikipedia usage estimates prevalence of influenza-like illness in the united states in near real-time. *PLoS computational biology*, **10**, e1003581. 6, 33

Mena, I., Nelson, M.I., Quezada-Monroy, F., Dutta, J., Cortes-Fernández, R., Lara-Puente, J.H., Castro-Peralta, F., Cunha, L.F., Trovão, N.S., Lozano-Dubernard, B. *et al.* (2016). Origins of the 2009 h1n1 influenza pandemic in swine in mexico. *Elife*, **5**, e16777. 3

Moat, H.S., Preis, T., Olivola, C.Y., Liu, C. & Chater, N. (2014). Using big data to predict collective behavior in the real world1. *Behavioral and Brain Sciences*, **37**, 92–93. 2

Moeller, J. (2015). A word on standardization in longitudinal studies: don't. *Frontiers in psychology*, **6**, 1389. 7

Molinari, N.A.M., Ortega-Sanchez, I.R., Messonnier, M.L., Thompson, W.W., Wortley, P.M., Weintraub, E. & Bridges, C.B. (2007). The annual impact of seasonal influenza in the us: measuring disease burden and costs. *Vaccine*, **25**, 5086–5096. 1

Montero, P., Vilar, J.A. *et al.* (2014). Tsclust: An r package for time series clustering. *Journal of Statistical Software*. 11

Moritz, S. & Bartz-Beielstein, T. (2017). imputeTS: Time Series Missing Value Imputation in R. *The R Journal*, **XX**, 1–12. 7

Nguyen, T., Henningsen, K.H., Brehaut, J.C., Hoe, E. & Wilson, K. (2011). Acceptance of a pandemic influenza vaccine: a systematic review of surveys of the general public. *Infection and drug resistance*, **4**, 197. 2

Panning, M., Eickmann, M., Landt, O., Monazahian, M., Olschläger, S., Baumgarte, S., Reischl, U., Wenzel, J., Niller, H., Günther, S. *et al.* (2009). Detection of influenza a (h1n1) v virus by real-time rt-pcr. *Euro surveillance: bulletin européen sur les maladies transmissibles= European communicable disease bulletin*, **14**. 2, 7

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **2**, 559–572. 8

Peppa, M., John Edmunds, W. & Funk, S. (2017). Disease severity determines health-seeking behaviour amongst individuals with influenza-like illness in an internet-based cohort. *BMC Infectious Diseases*, **17**, 238. 1

Pfaff, B. (2008). Var, svar and svec models: Implementation within R package vars. *Journal of Statistical Software*, **27**. 10

Prati, G., Pietrantoni, L. & Zani, B. (2011). A Social-Cognitive Model of Pandemic Influenza H1N1 Risk Perception and Recommended Behaviors in Italy. *Risk Analysis*, **31**, 645–656. 4

Quinn, S.C., Parmer, J., Freimuth, V.S., Hilyard, K.M., Musa, D. & Kim, K.H. (2013). Exploring communication, trust in government, and vaccination intention later in the 2009 H1N1 pandemic: results of a national survey. *Biosecur Bioterror*, **11**, 96–106. 64

Reintjes, R., Das, E., Klemm, C., Richardus, J.H., Ke??ler, V. & Ahmad, A. (2016). "Pandemic public health paradox": Time series analysis of the 2009/10 influenza A/H1N1 epidemiology, media attention, risk perception and public reactions in 5 European countries. *PLoS ONE*, **11**, 1–15. 2, 5, 7, 35

Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, **20**, 53–65. 12

Rubin, G.J., Amlôt, R., Page, L. & Wessely, S. (2009). Public perceptions, anxiety, and behaviour change in relation to the swine flu outbreak: cross sectional telephone survey. *BMJ (Clinical research ed.)*, **339**, b2651. 4, 64

Rubin, G.J., Potts, H.W.W. & Michie, S. (2011). Likely uptake of swine and seasonal flu vaccines among healthcare workers. A cross-sectional analysis of UK telephone survey data. *Vaccine*, **29**, 2421–2428. 65

Rudisill, C. (2013). How do we handle new health risks? Risk perception, optimism, and behaviors regarding the H1N1 virus. *Journal of Risk Research*, **16**, 959–980. 7, 65

Sardá-Espinosa, A. (2017). Comparing time-series clustering algorithms in r using the dtwclust package. *R package vignette*. 12

Savas, E. & Tanriverdi, D. (2010). Knowledge, attitudes and anxiety towards influenza A/H1N1 vaccination of healthcare workers in Turkey. *BMC infectious diseases*, **10**, 281. 64

Seale, H., Heywood, A.E., McLaws, M.L., Ward, K.F., Lowbridge, C.P., Van, D. & MacIntyre, C.R. (2010). Why do I need it? I am not at risk! Public perceptions towards the pandemic (H1N1) 2009 vaccine. *BMC Infect Dis*, **10**, 99. 65

Seybert, H. & Lööf, A. (2010). Internet usage in 2010 – Households and Individuals. *Eurostat Data in focus*, 1–8. 2

Shaman, J. & Kohn, M. (2009). Absolute humidity modulates influenza survival, transmission, and seasonality. *Proceedings of the National Academy of Sciences*, **106**, 3243–3248. 3

Shapiro, S.S. & Wilk, M.B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, **52**, 591–611. 9

Sharpe, J.D., Hopkins, R.S., Cook, R.L. & Striley, C.W. (2016). Evaluating Google, Twitter, and Wikipedia as Tools for Influenza Surveillance Using Bayesian Change Point Analysis: A Comparative Analysis. *JMIR Public Health and Surveillance*, **2**, e161. 1

Signorini, A., Segre, A.M. & Polgreen, P.M. (2011). The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *PLoS ONE*, **6**. 34

Simonsen, L., Spreeuwenberg, P., Lustig, R., Taylor, R.J., Fleming, D.M., Kroneman, M., Van Kerkhove, M.D., Mounts, A.W., Paget, W.J., Echenique, H., Savy, V., Muscatello, D., MacIntyre, C.R., Dwyer, D.E., Azziz-Baumgartner, E., Homaira, N., Moura, F.E.A., Schuck, C., Akwar, H., Schanzer, D., Fuentes, R., Olea, A., Sotomayor, V., Feng, L., Yu, H., Mazick, A., M??lbak, K., Nielsen, J., Carrat, F., Lemaitre, M., Buchholz, U., Schweiger, B., H??hle, M., Vesenbeckh, S., Cowling, B., Leung, G., Tsang, T., Chuang, S.K., Bromberg, M., Kaufman, Z., Sugaya, N., Oka Ezoe, K., Hayashi, S., Matsuda, M., Lopez-Gatell, H., Alpuche-Aranda, C., Noyola, D., Chowell, G., van Asten, L., Meijer, A., van den Wijngaard, K., van der Sande, M., Baker, M., Zhang, J., Benavides, J.G., Munayco, C., Laguna-Torres, A., Rabczenko, D., Wojtyniak, B., Park, S.H., Lee, Y.K., Zolotusca, L., Popovici, O., Popescu, R., Ang, L.W., Cutter, J., Lin, R., Ma, S., Chen, M., Lee, V.J., Prosenc, K., Socan, M., Cohen, C., Larrauri, A., de Mateo, S., M??ndez, L.S., Sanz, C.D., Andrews, N., Green, H.K., Pebody, R., Saei, A., Shay, D. & Viboud, C. (2013). Global Mortality Estimates for the 2009 Influenza Pandemic from the GLaMOR Project: A Modeling Study. *PLoS Medicine*, **10**. 3

Singer, E. & Endreny, P.M. (1993). *Reporting on Risk: How the Mass Media Portray Accidents, Diseases, Other Hazards*. Russell Sage Foundation. 3

Sjöberg, L. (1998). Worry and risk perception. *Risk analysis*, **18**, 85–93. 4

Smith, K.C., Rimal, R.N., Sandberg, H., Storey, J.D., Lagasse, L., Maulsby, C., Rhoades, E., Barnett, D.J., Omer, S.B. & Links, J.M. (2013). Understanding newsworthiness of an emerging pandemic: International newspaper coverage of the h1n1 outbreak. *Influenza and other respiratory viruses*, **7**, 847–853. 3, 33, 35

Snedecor, G. & Cochran, W. (1989). Analysis of variance: the random effects model. *Statistical Methods. Iowa State University Press, Ames, IA*, 237–252. 8

Stephens-Davidowitz, S. (2014). The cost of racial animus on a black candidate: Evidence using google search data. *Journal of Public Economics*, **118**, 26–40. 2

Stephens-Davidowitz, S. & Varian, H. (2014). A hands-on guide to google data. *Tech. Rep.*. 6, 35

Sypsa, V., Livanios, T., Psichogiou, M., Malliori, M., Tsiodras, S., Nikolakopoulos, I. & Hatzakis, A. (2009). Public perceptions in relation to intention to receive pandemic influenza vaccination in a random population sample: evidence from a cross-sectional telephone survey. *Euro surveillance : bulletin européen sur les maladies transmissibles = European communicable disease bulletin*, **14**, 1–5. 65

Taylor, M., Raphael, B., Barr, M., Agho, K., Stevens, G. & Jorm, L. (2009). Public health measures during an anticipated influenza pandemic: Factors influencing willingness to comply. *Risk Manag Healthc Policy*, **2**, 9–20. 4

Tooher, R., Collins, J.E., Street, J.M., Braunack-Mayer, A. & Marshall, H. (2013). Community knowledge, behaviours and attitudes about the 2009 H1N1 Influenza pandemic: A systematic review. *Influenza and other Respiratory Viruses*, **7**, 1316–1327. 2, 36

Van, D., McLaws, M.L., Crimmins, J., MacIntyre, C.R. & Seale, H. (2010). University life and pandemic influenza: Attitudes and intended behaviour of staff and students towards pandemic (h1n1) 2009. *BMC Public Health*, **10**, 130. 64

Vosen, S. & Schmidt, T. (2011). Forecasting private consumption: survey-based indicators vs. google trends. *Journal of Forecasting*, **30**, 565–578. 2

Walt, S.v.d., Colbert, S.C. & Varoquaux, G. (2011). The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, **13**, 22–30. 8

Walterdrkide, D.W., Böhmer, M.M., Reiter, S., Krause, G. & Wichmann, O. (2012). Risk perception and information-seeking behaviour during the 2009 / 10 influenza A (H1N1) pdm09 pandemic in Germany. 1–8. 4

Ward Jr, J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, **58**, 236–244. 11

White, L.F. & Pagano, M. (2010). Reporting errors in infectious disease outbreaks, with an application to pandemic influenza a/h1n1. *Epidemiologic Perspectives & Innovations*, **7**, 12. 35

Pandemic (h1n1) 2009 - update 112. 3

Wikimedia traffic analysis report - page views per wikipedia language - breakdown. 6

Won, M., Marques-Pita, M., Louro, C. & Gonçalves-Sá, J. (2017). Early and Real-Time Detection of Seasonal Influenza Onset. *PLOS Computational Biology*, **13**, e1005330. 1, 2, 33

Yoko, I., G.B., C., L.A., M., M., L. & A.P., G. (2010). The dynamics of risk perceptions and precautionary behavior in response to 2009 (H1N1) pandemic influenza. *BMC Infectious Diseases*, **10**. 65

Appendix A: Supplementary Figures

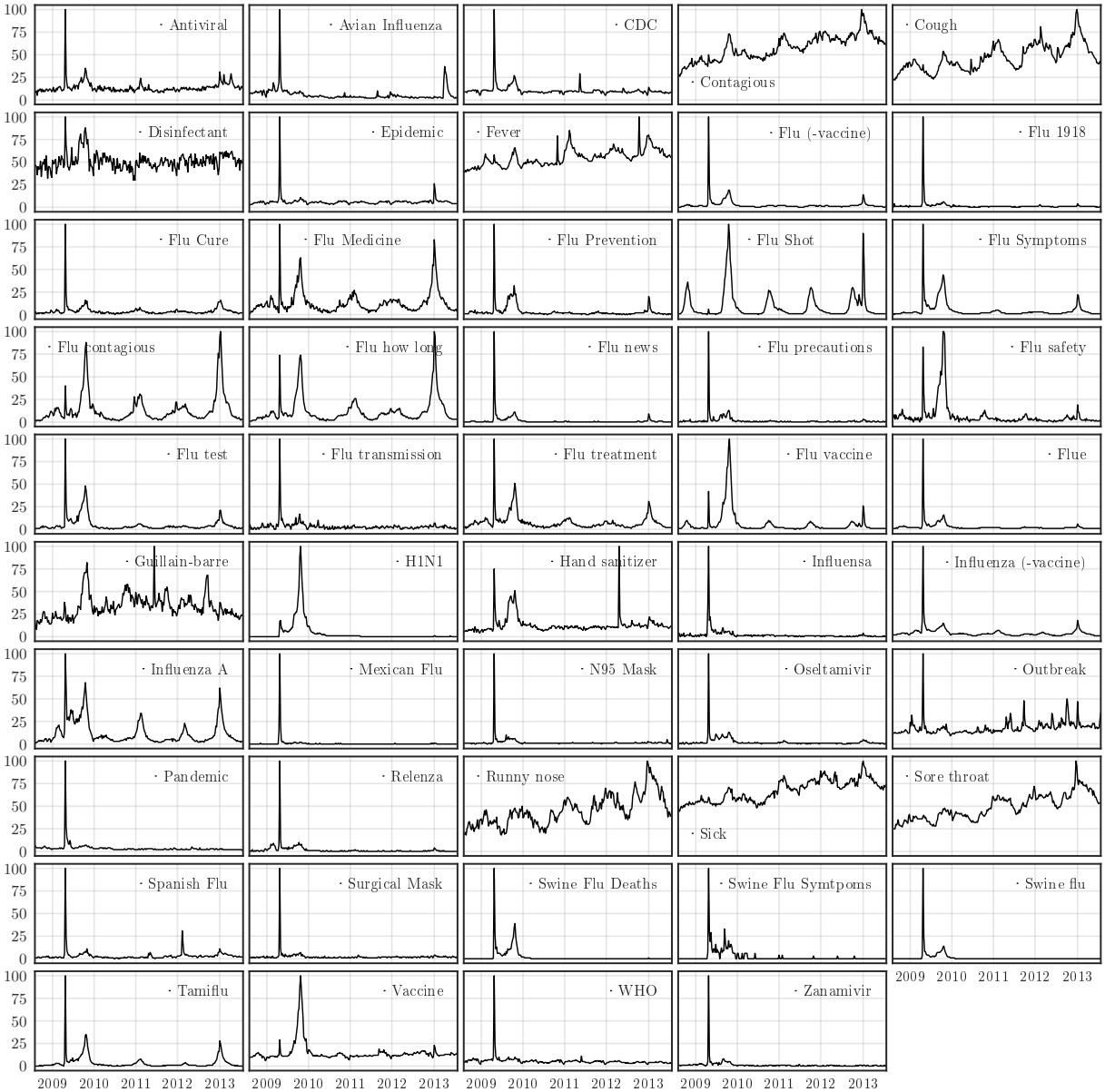

Figure S1: Google Trends - United States weekly series from July 2008 to July 2013. Y-axis: Google Search Volume Index (SVI).Referenced in page 15

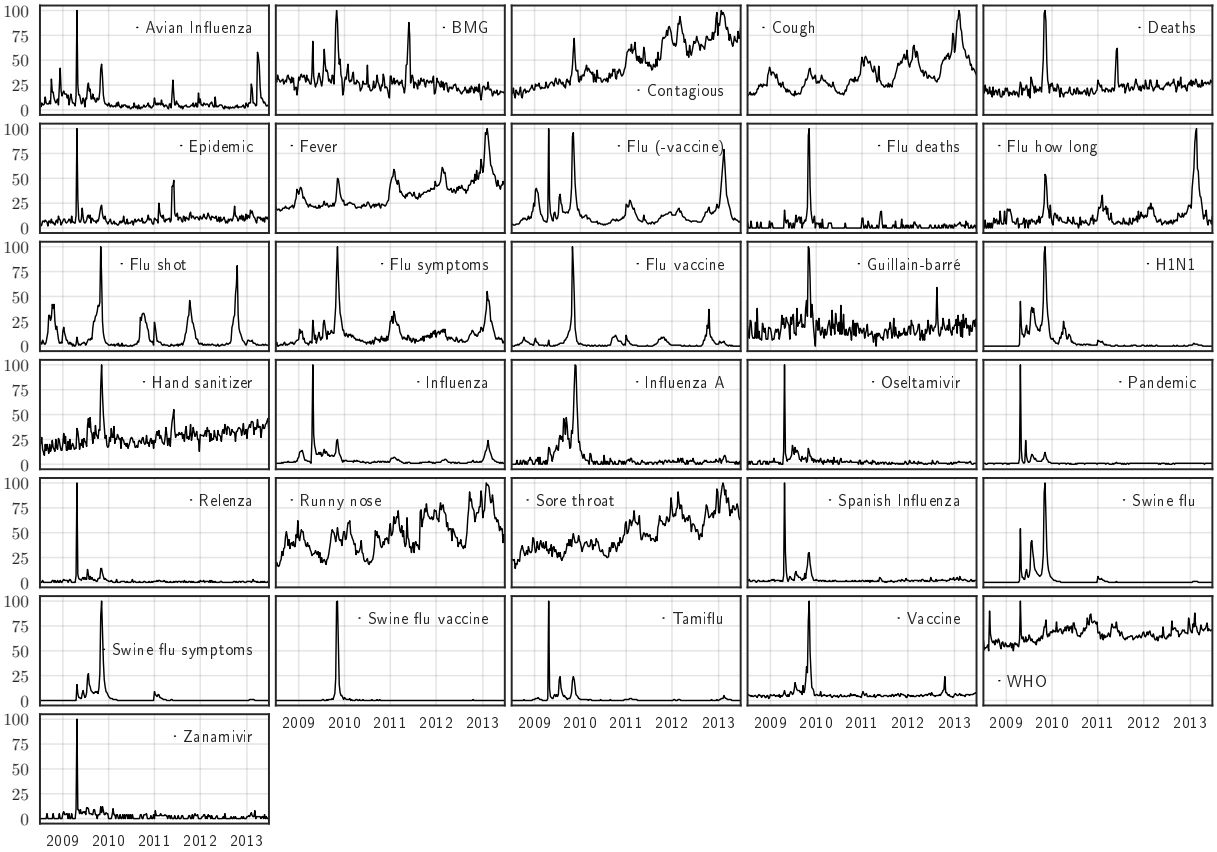

Figure S2: Google Trends - Germany. Weekly series from July 2008 to July 2013. Y-axis: Google Search Volume Index (SVI). Referenced in page 15.

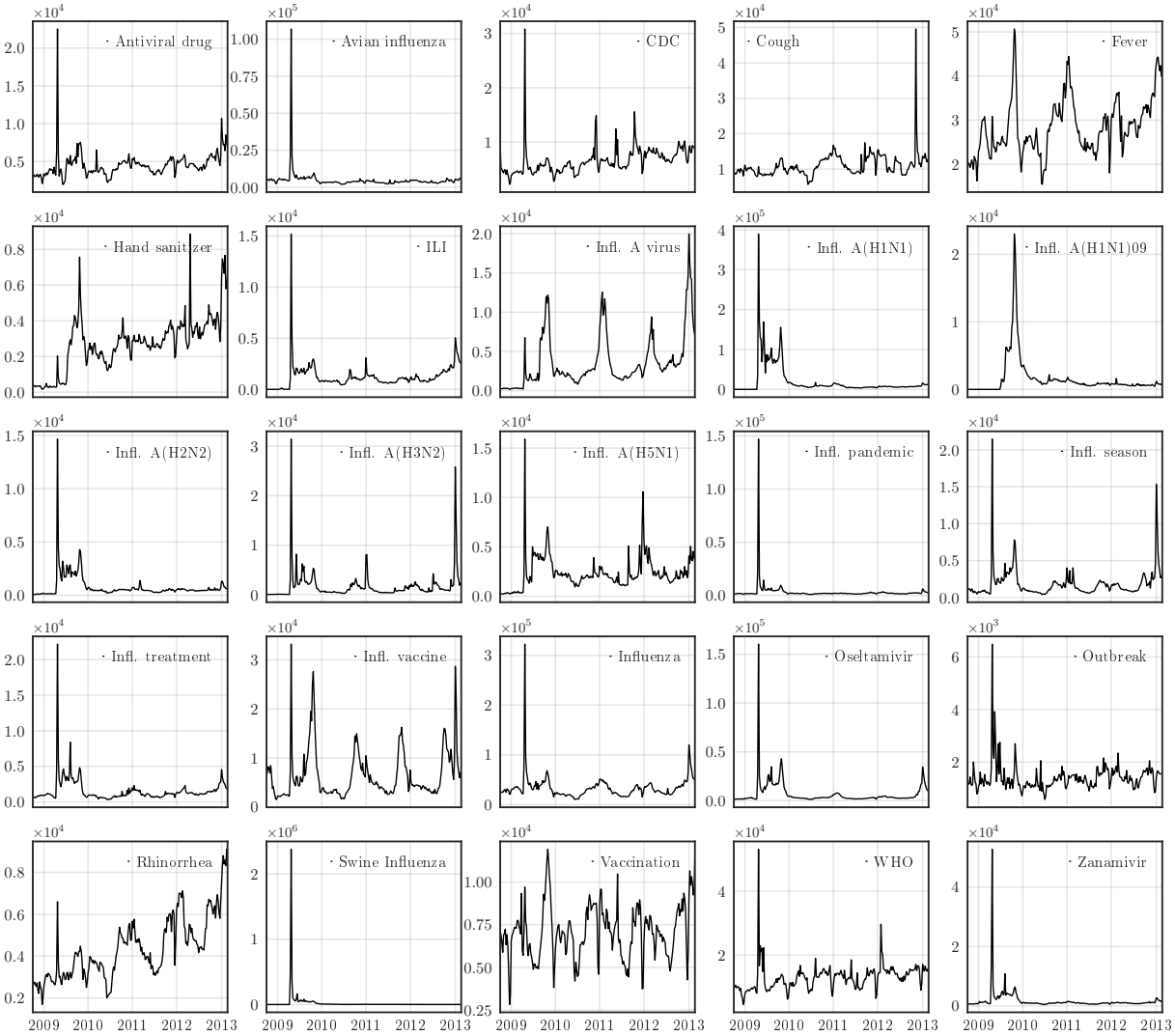

Figure S3: English Wikipedia. Weekly series, from July 2008 to July 2013. Y-axis is the absolute page-views count. Referenced in page 15.

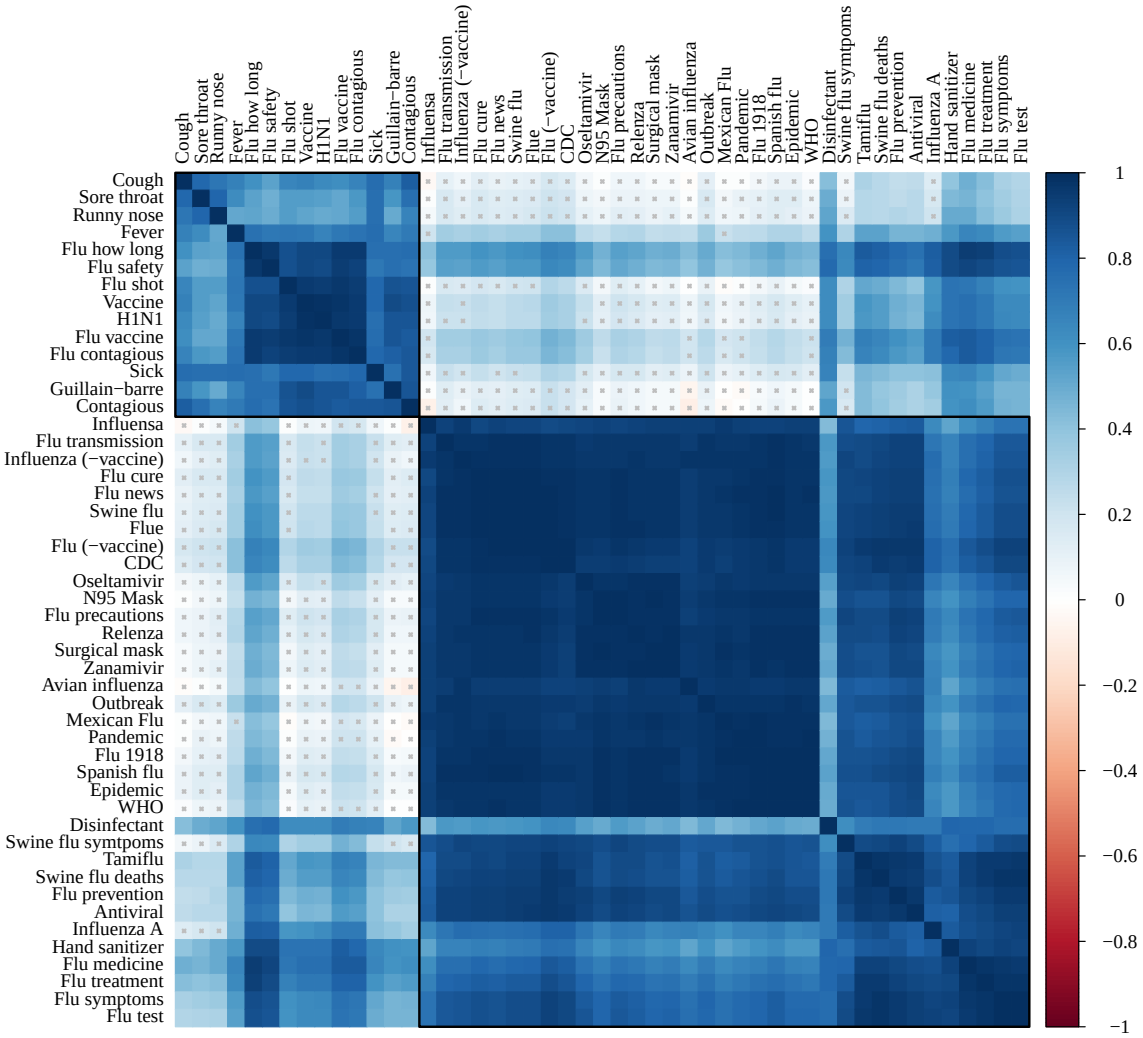

Figure S4: Correlation matrix of Google SVI (US). Non significant ($\alpha > 0.05$) correlations are indicated by a grey dot. Labels are ordered by hierarchical clustering using Pearson's correlation and Ward's linkage.

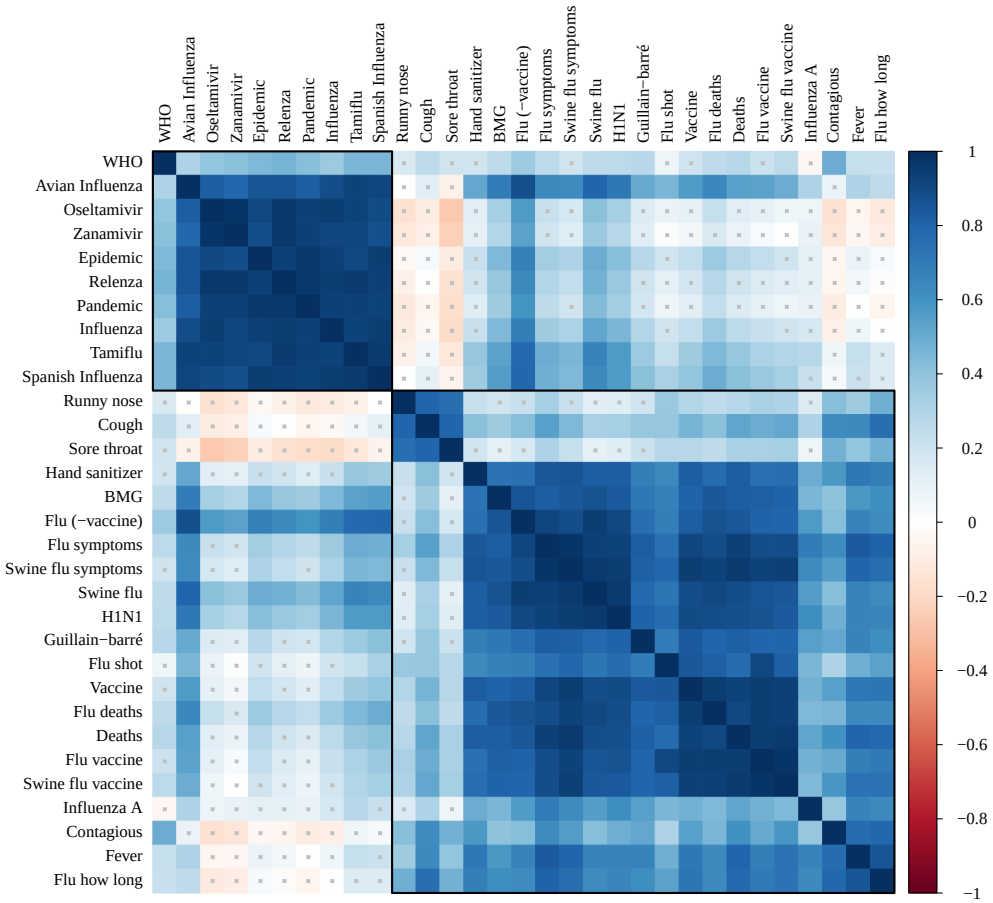

Figure S5: Correlation matrix of Google SVI (DE). Non significant ($\alpha > 0.05$) correlations are indicated by a grey dot. Labels are ordered by hierarchical clustering using Pearson's distance and Ward's linkage. Rectangles denote each cluster.

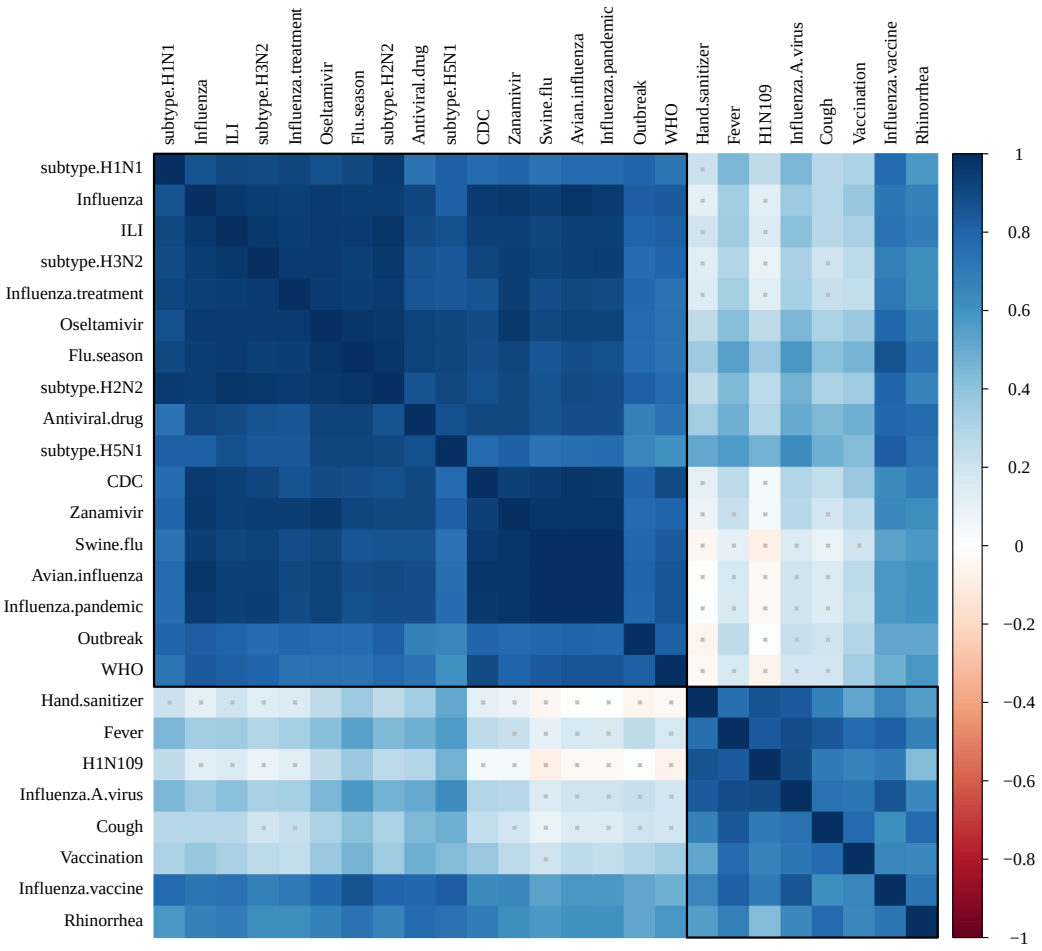

Figure S6: Correlation matrix of Wikipedia articles views. Non significant ($\alpha > 0.05$) correlations are indicated by a grey dot. Labels are ordered by hierarchical clustering using Pearson's distance and Ward's linkage.

Figure S7: Scatterplot, GT-US and news counts. * denotes $p - value < 0.05$, ** denotes p-value<0.001

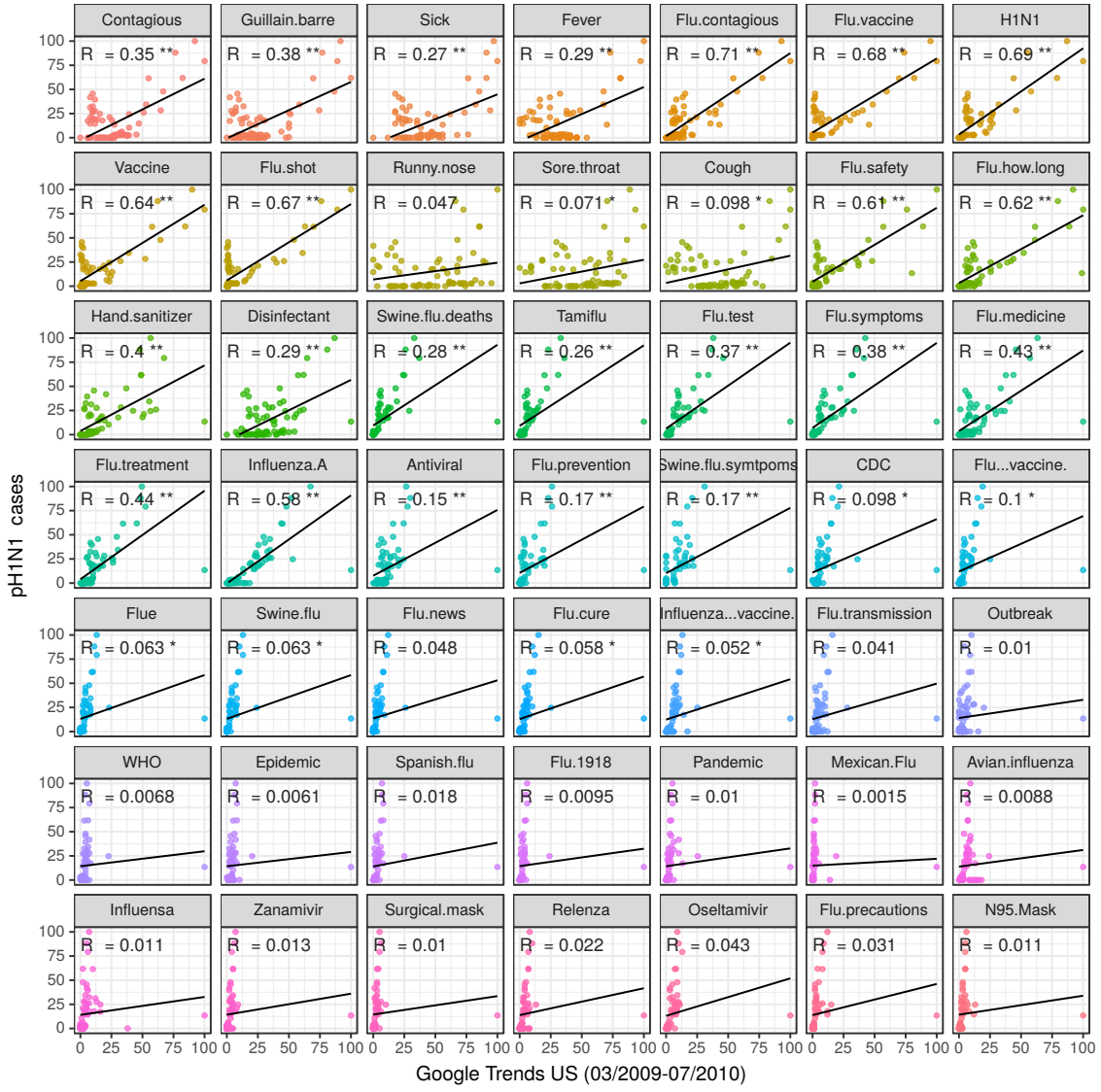

Figure S8: Scatterplot, GT-US and pH1N1 cases. * denotes $p - value < 0.05$, ** denotes p-value<0.001

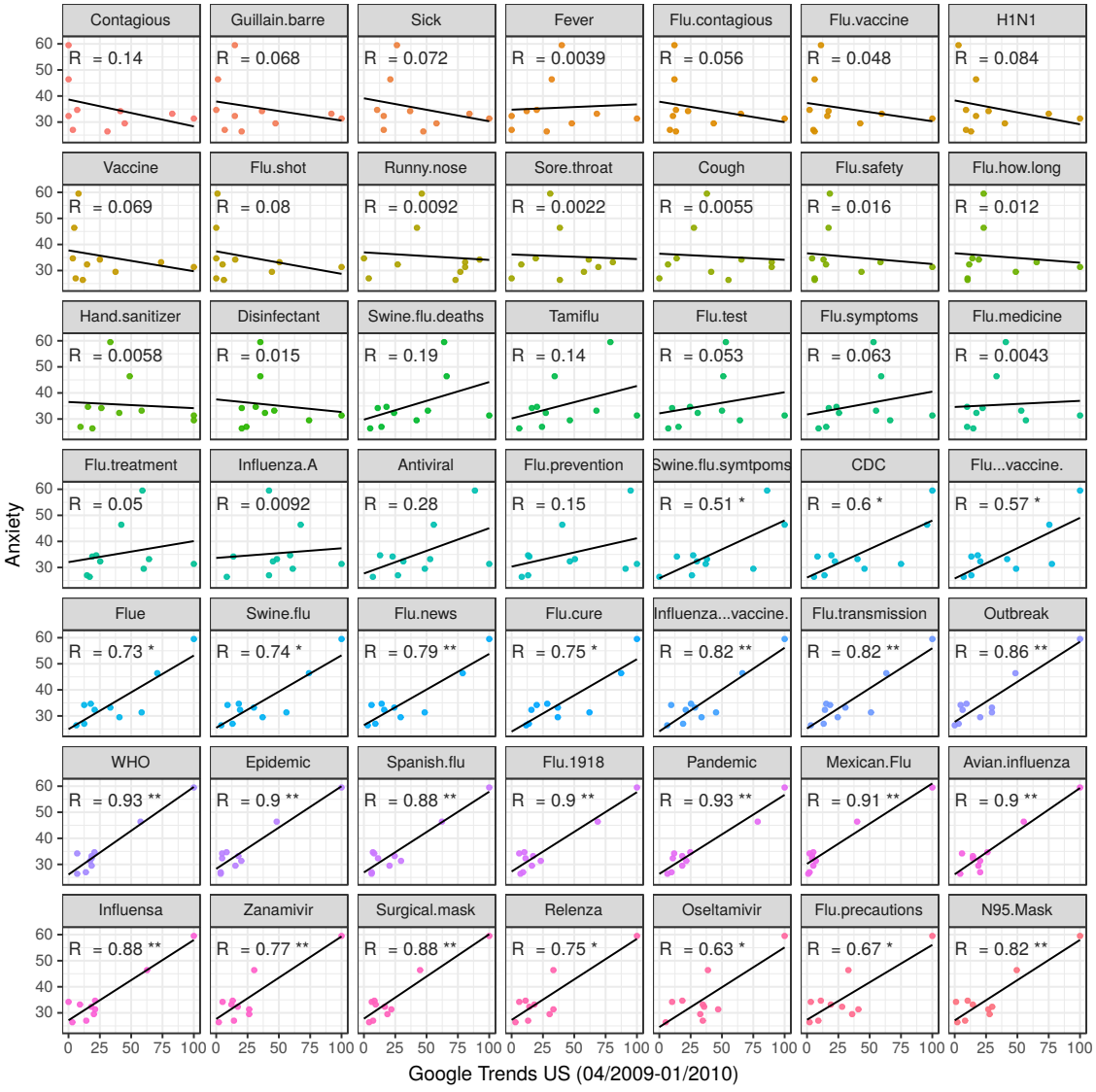

Figure S9: Scatterplot, GT-US and Anxiety levels. * denotes p-value<0.05, ** denotes p-value<0.001

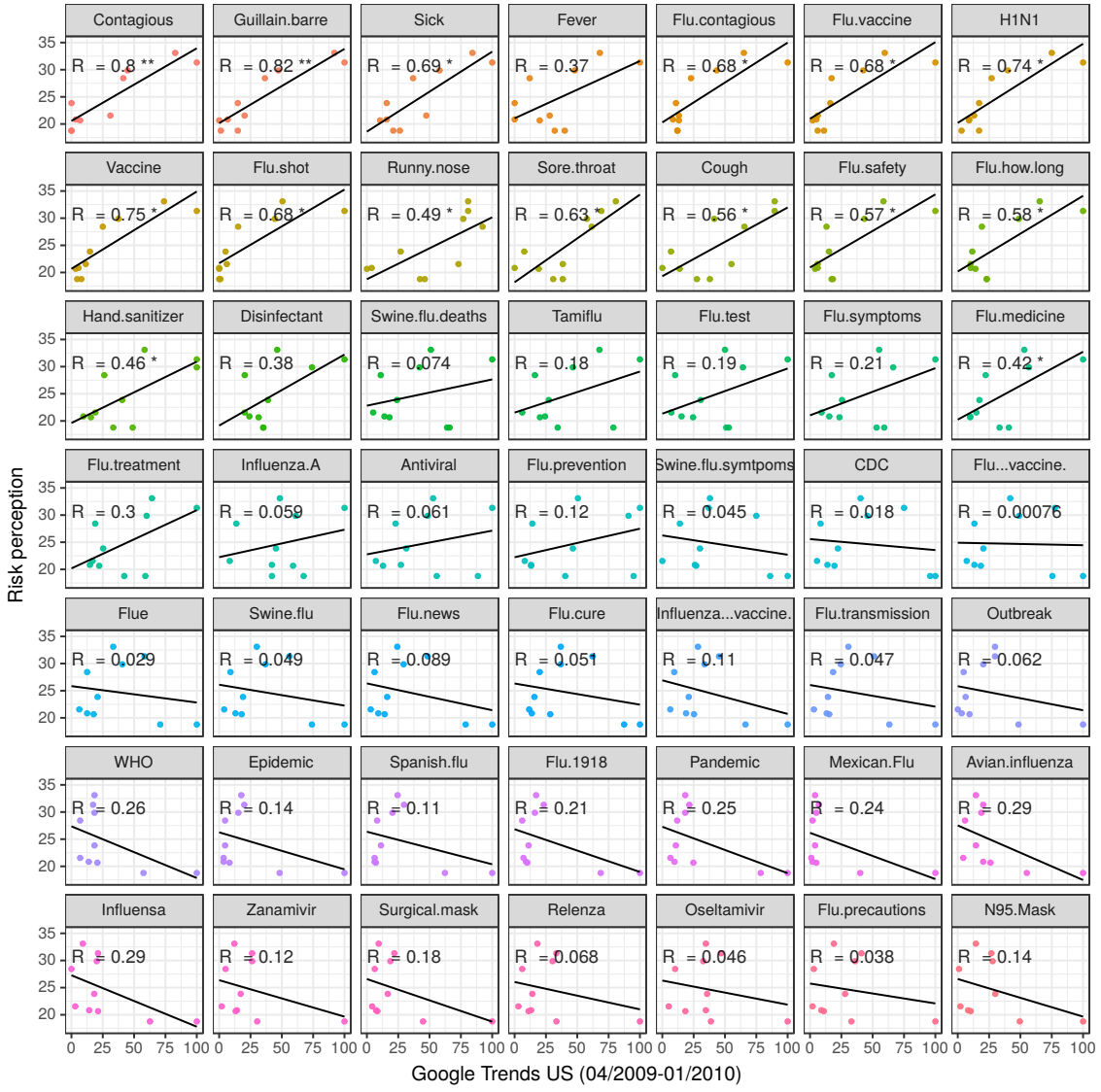

Figure S10: Scatterplot, GT-US and Risk perception levels. * denotes p-value<0.05, ** denotes p-value<0.001

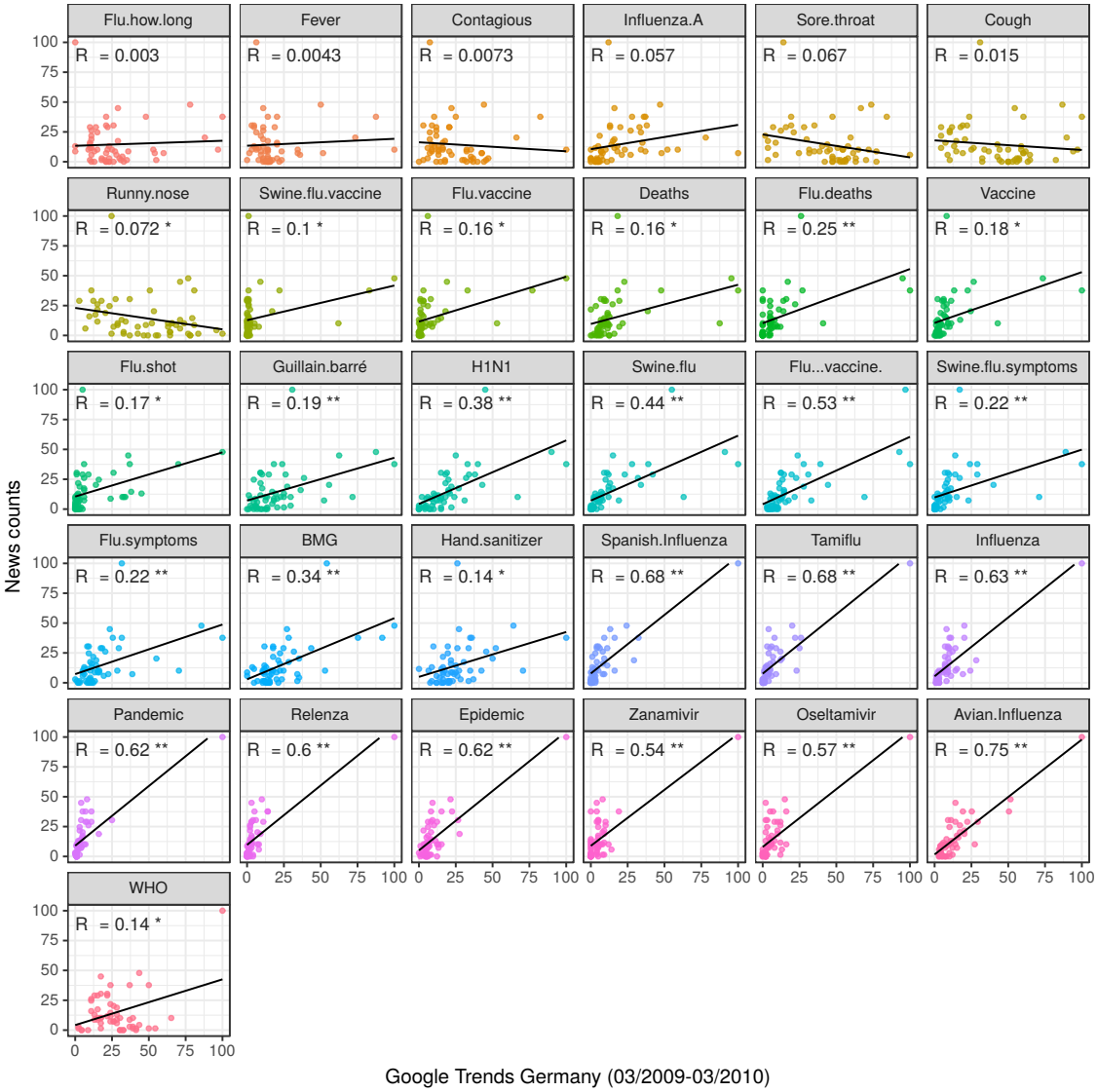

Figure S11: Scatterplot, GT-DE and news counts. * denotes p-value<0.05, ** denotes p-value<0.001

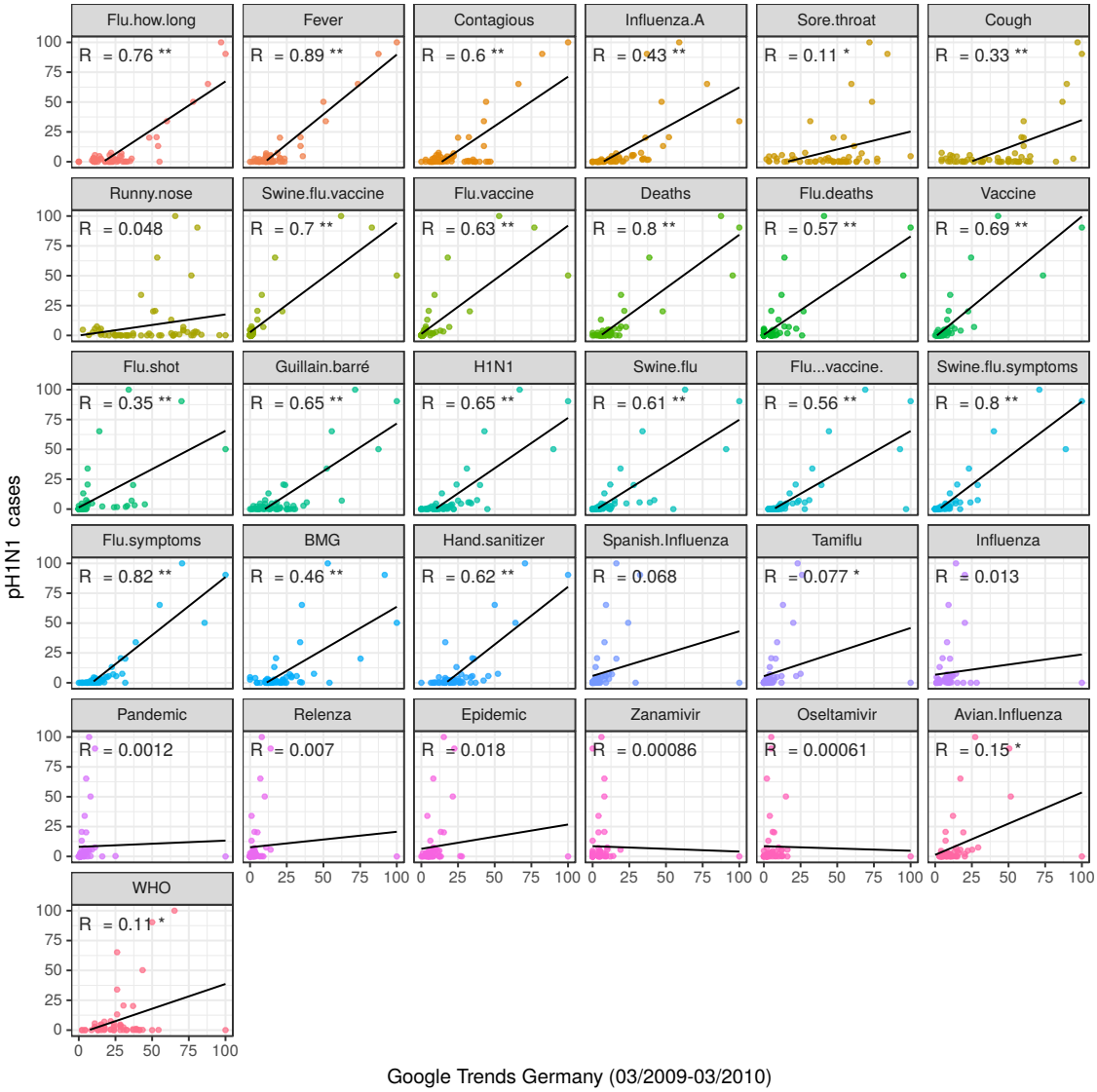

Figure S12: Scatterplot, GT-DE and pH1N1 cases. * denotes p-value$<$0.05, ** denotes p-value$<$0.001

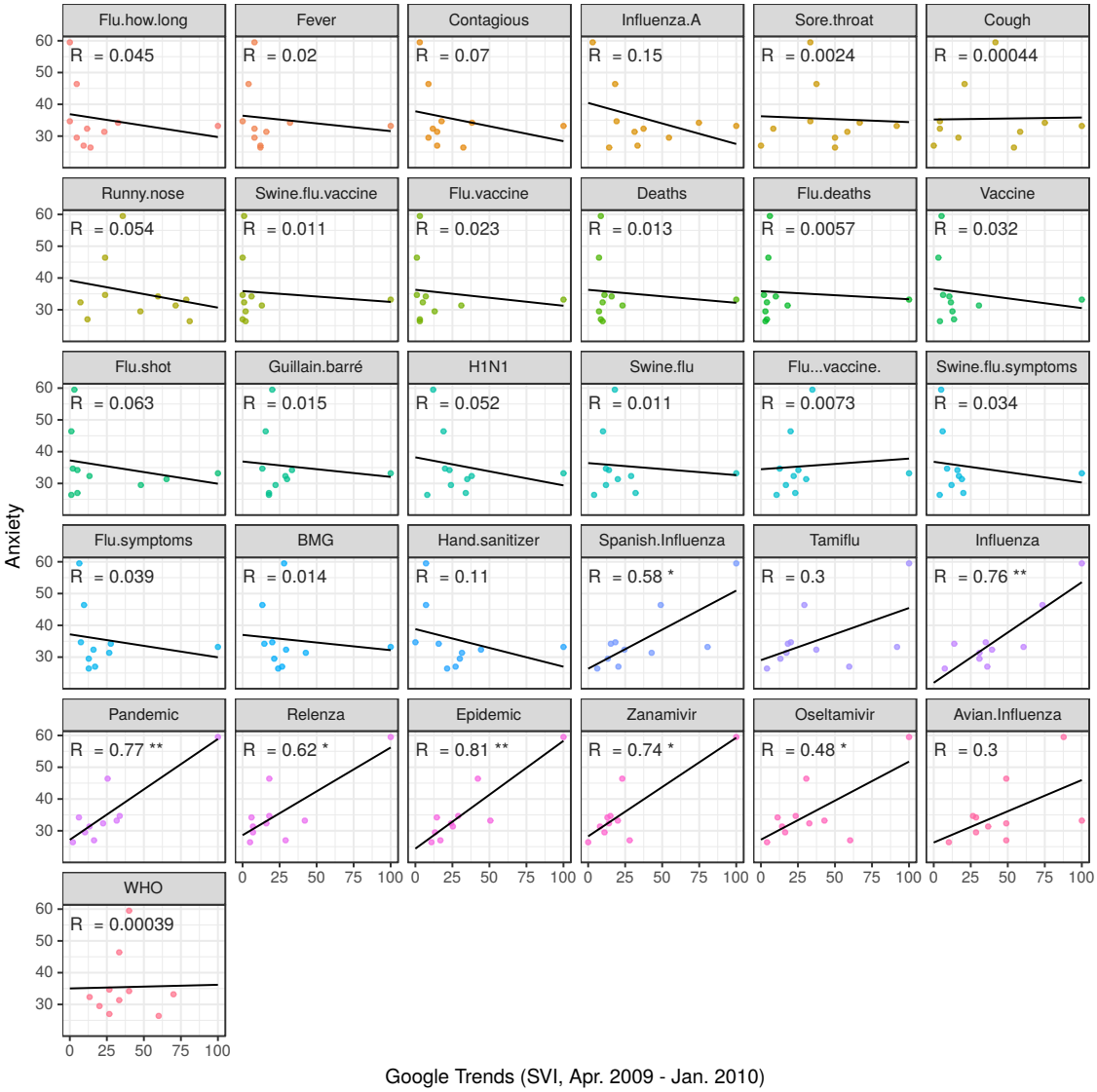

Figure S13: Scatterplot, GT-DE and Anxiety levels. * denotes p-value<0.05, ** denotes p-value<0.001

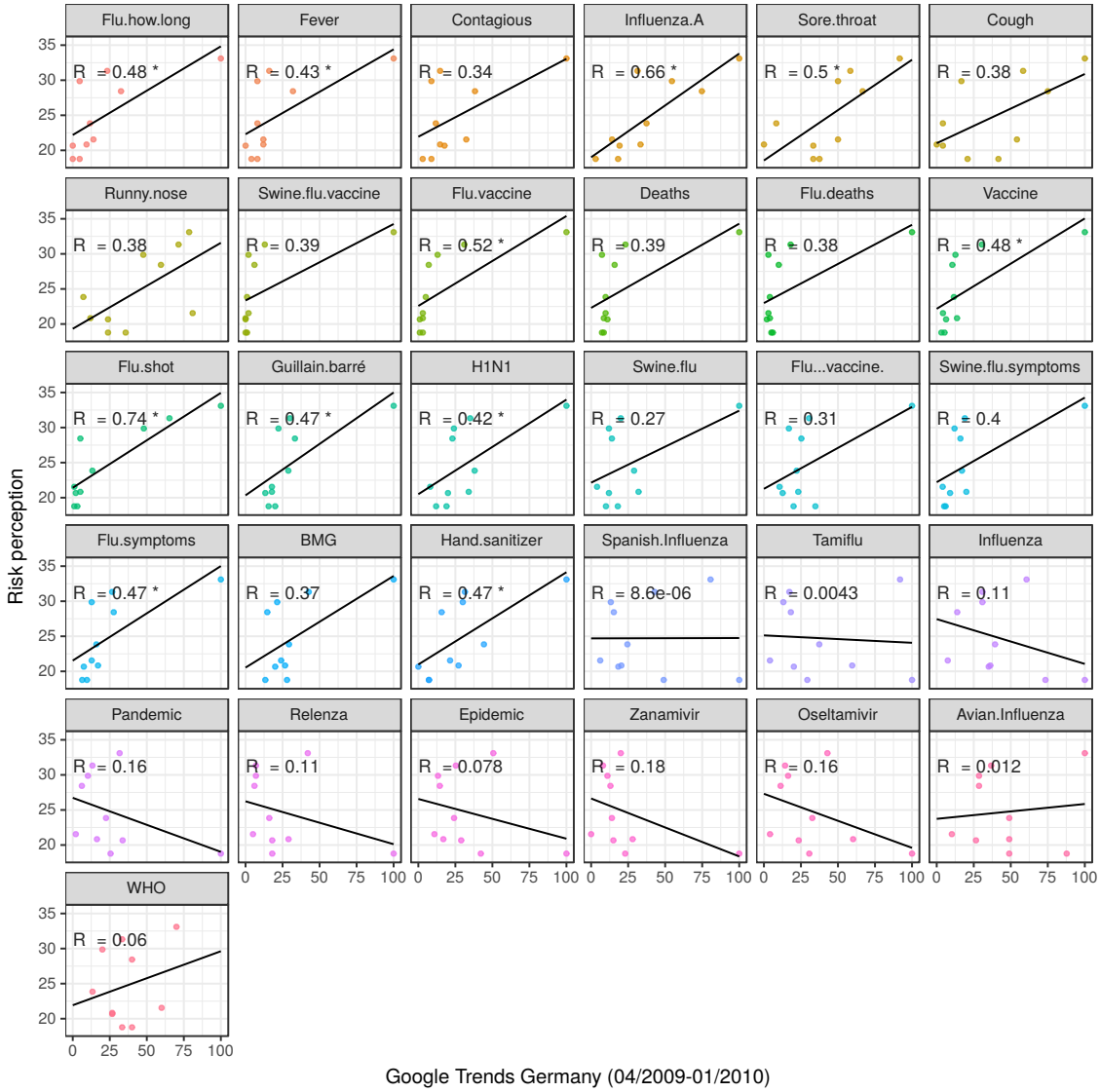

Figure S14: Scatterplot, GT-DE and Risk perception levels. * denotes p-value<0.05, ** denotes p-value<0.001

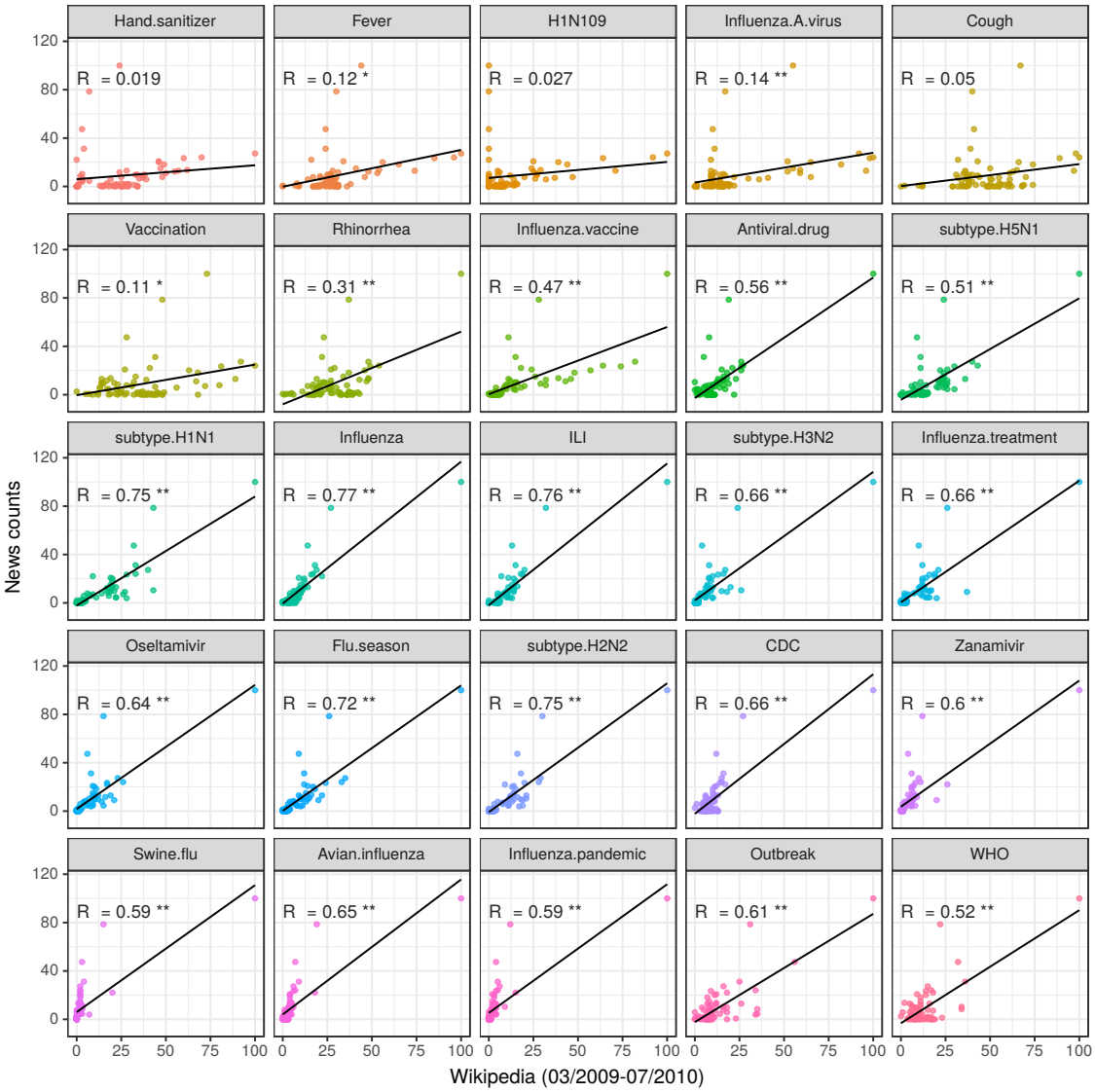

Figure S15: Scatterplot of Wikipedia page-views and news counts. * denotes p-value<0.05, ** denotes p-value<0.001

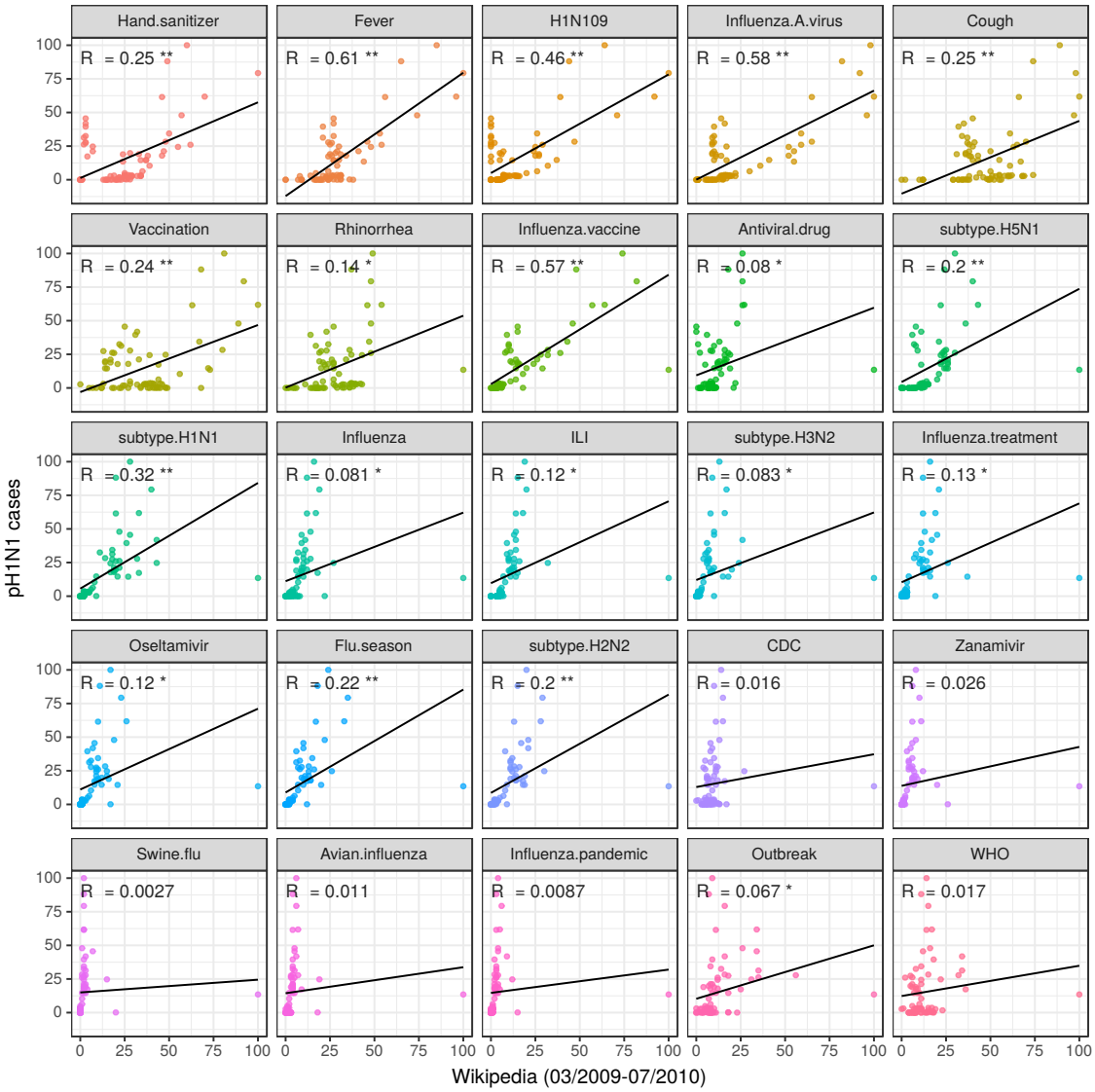

Figure S16: Scatterplot of Wikipedia page-views and pH1N1 cases. * denotes p-value$<$0.05, ** denotes p-value$<$0.001

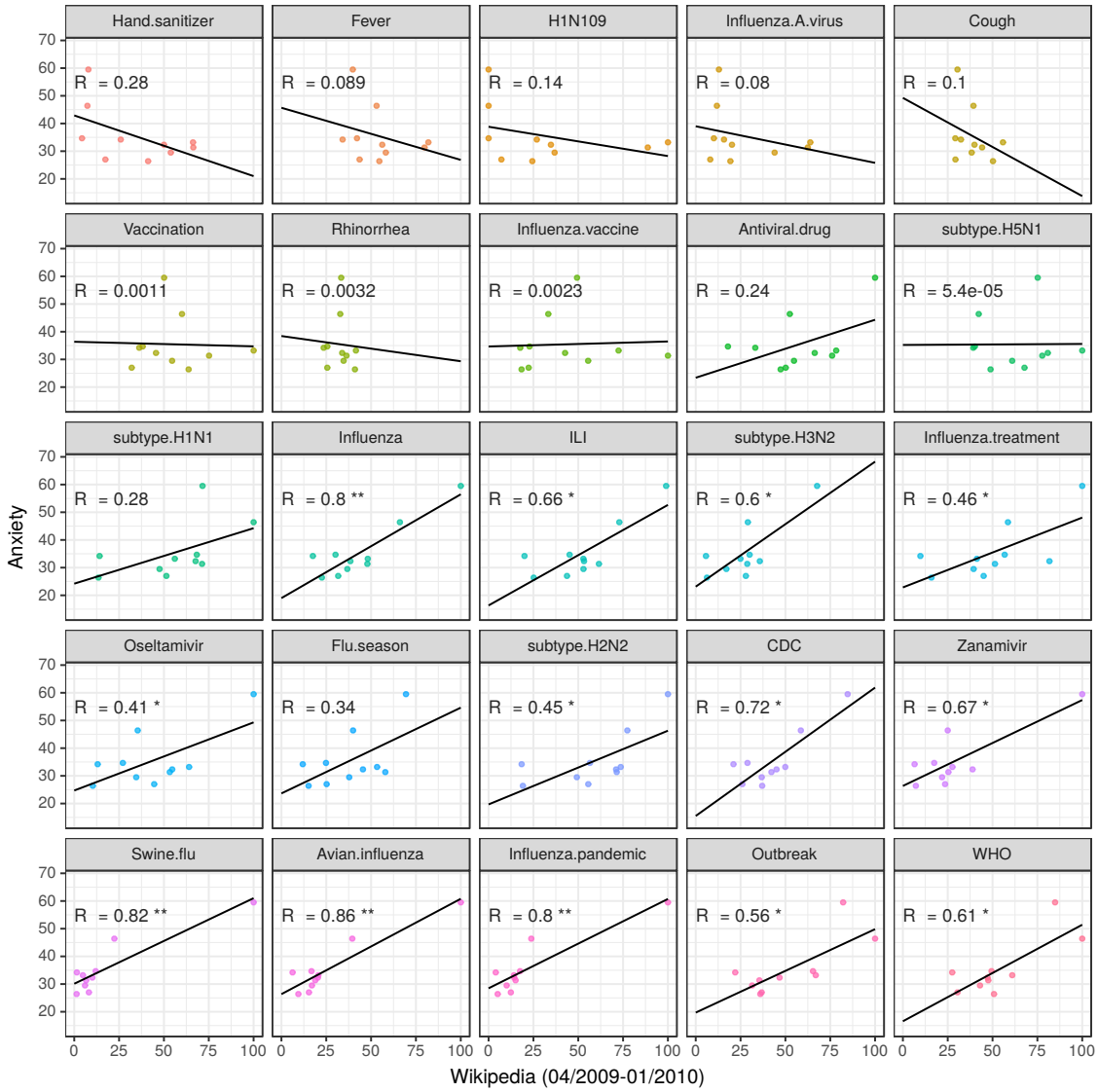

Figure S17: Scatterplot of Wikipedia page-views and Anxiety levels. * denotes p-value<0.05, ** denotes p-value<0.001

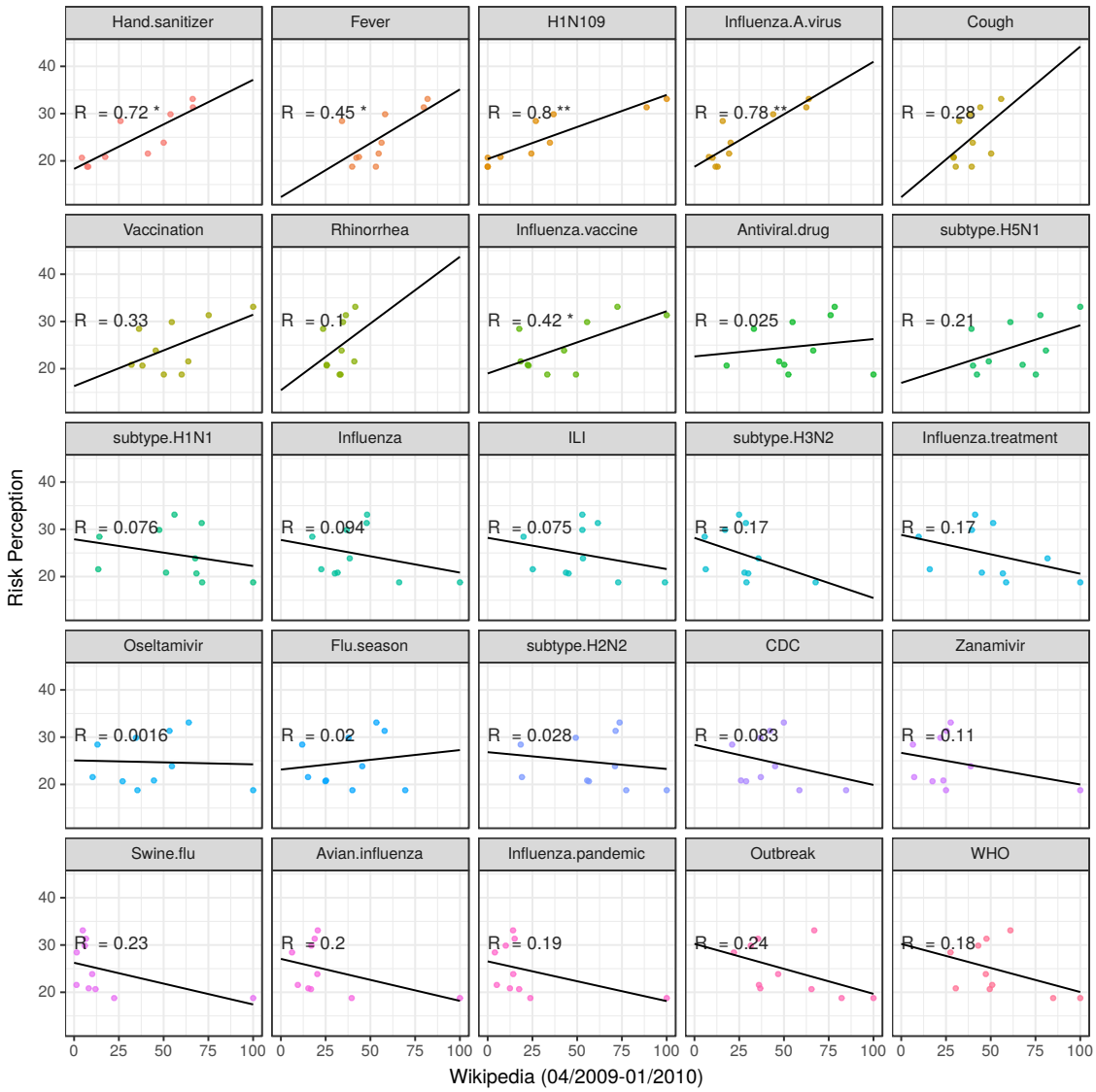

Figure S18: Scatterplot, Wikipedia page-views and Risk perception levels. * denotes p-value<0.05, ** denotes p-value<0.001

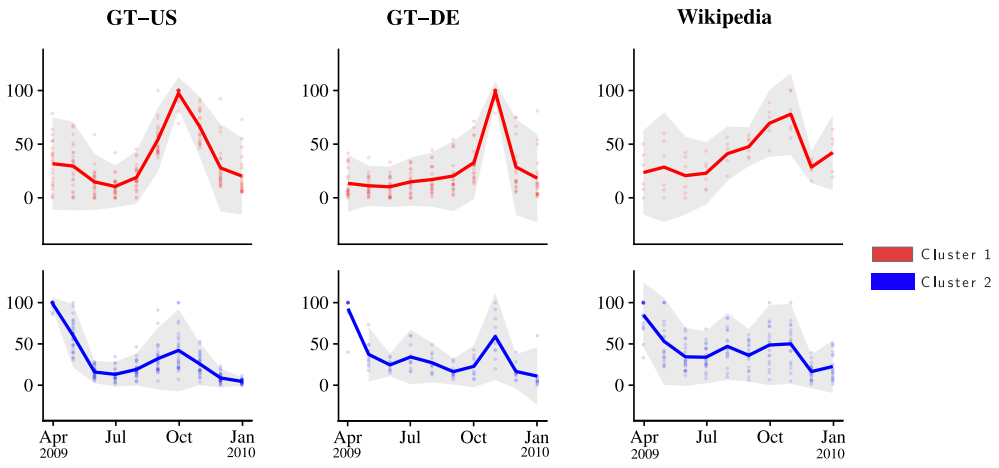

Figure S19: Centroid and standard deviation (grey shade) of Cluster 1 and Cluster 2 of Google Trends (US, DE) and Wiki-EN, regarding monthly data from April 2009 to January 2010. GT-US C1 $n = 26$, C2 $n = 23$; GT-DE C1 $n = 21$, C2 $n = 10$; Wiki-EN C1 $n = 6$, C2 $n = 19$. Referenced in page page 28.

Appendix B: Supplementary Tables

B.1 Collected terms

GT-US

Contagious, Guillain-barre, Sick, Fever, Flu contagious, Flu vaccine, H1N1, Vaccine, Flu shot, Runny nose, Sore throat, Cough, Flu safety, Flu how long, Hand sanitizer, Disinfectant, Swine flu deaths, Tamiflu, Flu test, Flu symptoms, Flu medicine, Flu treatment, Influenza A, Antiviral, Flu prevention, Swine flu symptoms, Flu news, Flu cure, Influenza (-vaccine), Flu transmission, Outbreak, WHO, Epidemic, Spanish Flu, Flu 1918, Pandemic, Mexican Flu, Avian influenza, Influensa, Zanamivir, Surgical mask, Relenza, Oseltamivir, Flu precautions, N95 mask

GT-DE

Schweinegrippe (Swine flu), Impfstoff (Vaccine), Grippe -impfstoff (Flu -vaccine), Grippeimpfung (Flu vaccine), Grippeschutzimpfung (Flu shot), Schweine grippeimpfung (Swine flu vaccine), Influenza, Guillain barre, Tamiflu, Relenza, Oseltamivir, Zanamivir, Grippe Symptome (Flu symptoms), Spanische grippe (Spanish flu), Schweinegrippe Symptome (Swine flu symptoms), Fieber (Fever), Husten (Cough), schnupfen (Runny nose), Halsschmerzen (Sore throat), ansteckend (Contagious), pandemie (Pandemic), epidemie (Epidemic), Vogelgrippe (Avian flu), H1N1, BMG, Wie lange grippe (flu how long), desinfektionsmittel (hand sanitizer), Grippe Todesfälle (flu deaths), Todesfälle (deaths), grippe a (influenza A), WHO

Wiki-EN

World Health Organization, Outbreak, Influenza pandemic, Avian influenza, Swine flu, Zanamivir, CDC, Influenza subtype H2N2, Flu season, Influenza treatment, Influenza subtype H3N2, ILI, Influenza, Influenza subtype H1N1, Influenza subtype H5N1, Antiviral drug, Influenza vaccine, Rhinorrhea (runny nose), Vaccination, Cough, Influenza A Virus, Influenza subtype H1N109, Fever, Hand sanitizer

B.2 Cluster validation index

Table T1: Google Trends (US) Cluster validation indexes. Higher Sil, Dunn and lower DB* indicate optimal partition.

	$N_c = 2$	$N_c = 3$	$N_c = 4$	$N_c = 5$
Sil	**0.50**	0.47	0.49	0.49
Dunn	0.20	0.26	**0.31**	0.31
DB*	**0.81**	1.32	1.13	1.45

Table T2: Google Trends (DE) Cluster validation indexes. Higher Sil, Dunn and lower DB* indicate optimal partition.

	$N_c = 2$	$N_c = 3$	$N_c = 4$	$N_c = 5$
Sil	**0.67**	0.49	0.52	0.57
Dunn	0.14	0.14	**0.25**	0.13
DB*	**0.41**	1.13	0.75	0.93

Table T3: Wikipedia (EN) Cluster validation indexes.Higher Sil, Dunn and lower DB* indicate optimal partition.

	$N_c = 2$	$N_c = 3$	$N_c = 4$	$N_c = 5$
Sil	**0.54**	0.38	0.40	0.36
Dunn	**0.52**	0.38	0.50	0.32
DB*	**0.79**	1.35	1.23	1.44

Appendix C: Surveys Summary

Table T4: Summary of the collected surveys regarding Anxiety to pH1N1, considering the period from April 2009 to January 2010. UK=United Kingdom; NL=Netherlands; TK=Turkey; EU=Europe; IT=Italy; US=United States, AU=Australia, PT=Portugal

Article Method	Period	Country	Question / Index	%
Rubin *et al.* (2009)	May 2009	UK	Trait anxiety score	23.8
Bults *et al.* (2011)	April-June, August, 2009	NL	Worry about pH1N1 (worried-very worried)	61, 61, 40, 38
Gallup (2013)	December 2009	EU	Worry about pH1N1 (very, intermediate)	42
Savas & Tanriverdi (2010)	November 2009	TK	Trait anxiety score	40
Van *et al.* (2010)	July-September 2009	AU	Anxiety about pH1N1	36
Eastwood *et al.* (2010)	August-September 2009	AU	Not disclosed, "A variety of questions assessed anxiety"	22.2
Ferrante *et al.* (2011)	November/09-January/10	IT	Worry about pH1N1 (very, somewhat)	26.4
Harvard (2009)	May,June 2009	US		46, 41
Goodwin *et al.* (2009)	April-May 2009	EU	Concerned over pH1N1 (Very, somewhat)	58
Quinn *et al.* (2013)	May-June 2009	US	Concerned over pH1N1	46
GripeNet (2009)	June-July 2009	PT	Anxiety over pH1N1 (intermediate-very high)	18

Table T5: Summary of the collected surveys related to risk perception to pH1N1, considering the period from April 2009 to January 2010. UK=United Kingdom; NL=Netherlands; TK=Turkey; EU= Europe; IT=Italy; GR=Greece, ES=Spain, AU=Australia, PT=Portugal

Article Method	Period	Country	Question / Index	%
Jones & Salathé (2009)	April-May, 2009	US	Risk of infection (intermediate-highly likely)	22
Yoko *et al.* (2010)	April-May, 2009	US	Likelihood of contact with infected	25.6
Rudisill (2013)	October, 2009	UK	Risk of infection (0-100 scale)	35.7
Gidengil *et al.* (2012)	May-July 2009, September-November 2009, January 2010	US	Risk of infection (0-100 scale)	9,10,12,18,17,11
Bults *et al.* (2011)	April-June, August, 2009	NL	Susceptibility to pH1N1 (quite, very)	18,22,28
Sypsa *et al.* (2009)	August-October, 2009	GR	Likelihood of infection (very, quite)	27.4
Ferrante *et al.* (2011)	November/09-January/10	IT	Risk of infection (high)	33.1
Rubin *et al.* (2011)	September-October 2009	UK	Possibility of infection (very, fairly)	17.2
Seale *et al.* (2010)	September-October, 2009	AU	Risk of infection (very high, high)	45.3
Agüero *et al.* (2011)	December, 2009	ES	Likelihood of infection	15.7
Gallup (2013)	December, 2009	EU	Likelihood of infection (very, rather likely)	36.5
GripeNet (2009)	June-July 2009	PT	Likelihood of infection (very-intermediate)	29.5