Utility of Google Trends in anticipating Coronavirus Disease 2019
(COVID-19) outbreaks in Poland

Authors: Brandon Michael Henry, Ivan Szergyuk, Maria Helena Santos de Oliveira, Giuseppe Lippi, Grzegorz Juszczyk, Marcin Mikos

Article type: Research letter

Received: January 8, 2021.

Accepted: March 19, 2021.

Published online: March 26, 2021.

ISSN: 1897-9483
Utility of Google Trends in anticipating Coronavirus Disease 2019 (COVID-19) outbreaks in Poland

Short title: Google Trends and COVID-19 outbreaks

Brandon Michael Henry¹,*, Ivan Szergyuk²,*, Maria Helena Santos de Oliveira³, Giuseppe Lippi⁴, Grzegorz Juszczyk⁵, Marcin Mikos⁶

¹ Cardiac Intensive Care Unit, The Heart Institute, Cincinnati Children’s Hospital Medical Center, Ohio, USA
² Faculty of Medicine, Jagiellonian University Medical College, Krakow, Poland
³ Department of Statistics, Federal University of Parana, Curitiba, Brazil
⁴ Section of Clinical Biochemistry, University of Verona, Verona, Italy
⁵ Department of Public Health, Medical University of Warsaw, Warsaw, Poland
⁶ Bioinformatics and Public Health Department, Faculty of Medicine and Health Sciences, Andrzej Frycz Modrzewski Krakow University, Krakow, Poland

*These authors contributed equally to this work.

Conflict of interest: none declared.

Correspondence to: Brandon Michael Henry, MD

Cardiac Intensive Care Unit, The Heart Institute, Cincinnati Children's Hospital Medical Center, 3333 Burnet Ave., Cincinnati, OH 45229, USA.

Email: brandon.henry@cchmc.org

Phone/Fax: +1 716 598 86
Introduction

The majority of worldwide laboratories have reported considerable shortages in test kits, reagents, and qualified personnel required to perform severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) diagnostic testing which are likely to continue well into 2021 [1]. These shortages can lead to underestimations of the true epidemiological state of Coronavirus Disease 2019 (COVID-19) and point to the need for alternative surveillance methods to anticipate outbreaks and the course of the pandemic.

Internet user data, termed infodemiology metrics [2], have been widely employed to track the epidemiological trajectory of past infectious outbreaks, such as influenza [3]. The value of infodemiological data in Poland is promising, given that 66.7% of the Polish population use the internet for health-related queries [4]. Emerging evidence suggests that the volume of COVID-19 specific Google searches using symptom keywords correlates with the number of actual COVID-19 cases reported [5,6]. COVID-19 infodemiology may be aided by the presence of pathognomonic symptoms, in particular, loss of taste or smell. Recent observations have demonstrated 11-fold increased odds for having olfactory and/or gustatory dysfunction in COVID-19 compared to other causes of acute respiratory infection [7], underscoring the potential utility of monitoring these characteristic symptoms in the community to predict current COVID-19 cases and inform public health responses.

In this investigation, we aimed to determine whether Google search trends pertaining to anosmia and dysgeusia ("utrata węchu", "utrata smaku"), as well as other COVID-19 symptoms, correlate with the past surveillance data in Poland over the course of the pandemic.
Methods

Google relative search term interest values in Poland were gathered using Google Trends for the period between May 18th and November 22nd, 2020, and aggregated by week in a Microsoft Excel spreadsheet (Microsoft, Redmond, WA, United States), given that number of COVID-19 cases was also collected on a weekly basis to avoid daily variations in reporting. Each unit in Google Trends represents relative search interest per day on a 100-point scale, where the maximum is set by the highest search volume for a particular term in the period studied, and a weekly score may amount up to 700 points. This study focused on Polish Google search terms concerning loss of smell (“utrata węchu”), loss of taste (“utrata smaku”), as well as fever (“gorączka”), headache (“ból głowy”), cough (“kaszel”), dyspnea (“duszność”), and shortness of breath (“problemy z oddychaniem”). Data on the number of new weekly cases in Poland was obtained via official reports from the Polish National Institute of Public Health [8], and tabulated together with the Google Trend scores throughout the same period.

For each search term, cross-correlation functions were plotted to determine the time lags that provided the highest possible correlations between the Google Trends time series and the weekly COVID-19 case count time series, as well as corresponding $P$ values and 95% Confidence Intervals (95%CI). Time series linear regression was performed to estimate the quantitative effect of Google Trend score increases on subsequent increases in weekly COVID-19 cases, and to identify the time lag with the highest predictive value. Variable selection was based on results of cross correlation functions, considering lags with highest cross correlations. These results were obtained through R Software (version 4.0.2; R Foundation for Statistical Computing). We chose a $P$ value of $<0.05$ as the threshold of statistical significance. Model performance was assessed using adjusted $R^2$ values and
graphic analysis to see how closely infection rate predictions based on Google Trend scores resemble actual surveillance records.

This study was conducted in compliance with the Declaration of Helsinki, under the terms of relevant local legislation. The analysis was based on searches of unrestricted, publicly available databases; thus, no informed consent or ethical committee approvals were required.

Results

Weekly COVID-19 cases and Google Trend scores for each symptom searched are presented in Supplementary material, Table S1. The highest Google Trend value out of all searched symptoms was observed for cough during the week of October 26th, and the lowest was observed for smell loss during the week of June 1st. The weekly number of new COVID-19 cases ranged from 1927 on the last week of June to 172,658 on the week of November 9th.

Results of cross correlation analysis can be found in Supplementary material, Table S2. Among symptoms searched, both smell loss and taste loss had the highest peak cross-correlations to the weekly case count time series, particularly at a one-week lag (r=0.91, 95%CI: 0.72-1, P<0.001 and r=0.93, 95%CI: 0.74-1, P<0.001, respectively). Figure 1A illustrates the resemblance of the pattern in Google Trend scores for loss of smell and taste and new weekly COVID-19 cases in the following weeks. The remaining symptoms had weaker cross associations at their respective optimal time lags (r between 0.41-0.85).

“Dyspnea” and ‘shortness of breath’ had the highest cross correlation coefficients at no time lag (r=0.58, 95%CI: 0.39-0.77, P=0.003 and r=0.41, 95%CI: 0.22-0.6, P=0.03, respectively). The term “fever” required a 2-week lag for optimal correlation (r=0.75, 95%CI: 0.56-0.94, P<0.001), and both headache and cough required a 1-week lag (r=0.85, 95%CI: 0.66-1, P<0.001 and r=0.75, 95%CI: 0.57-0.95, P<0.001, respectively). Figure 1B shows trends in
Google searches for non-smell and non-taste related symptoms, with patterns clearly less similar to surveillance data.

The results of time series linear regression for smell loss and taste loss are shown in Supplementary material, Table S3. For smell loss, the effect of the two-week lagged time series was highly significant, and estimated at 340.7 (95%CI: 327.2-354.2; \( P<0.001 \)), meaning the number of predicted COVID-19 cases were increased by 340.7 per one-unit increase in the Google Trends value for smell loss two weeks before. The effect of the one-week lagged time series for smell loss was not significant. This linear regression model had an adjusted \( R^2 \) of 0.99. For taste loss, both the one-week and the two-week lagged time series were significant for estimating the weekly number of COVID-19 cases. The model expression indicates that the number of predicted COVID-19 cases rose by 111.5 per one-unit increase in the Google Trends value for taste loss one week before (95%CI: 65.4-157.6; \( P<0.001 \)), and by 251.2 per one unit increase two weeks prior (95%CI: 201.2-301.1; \( P<0.001 \)). The adjusted \( R^2 \) for this model was 0.98. The fitted values of these linear regression models for smell loss and taste loss, as well as the trend in COVID-19 cases reported, are represented in Figure 1C and 1D, respectively.

**Discussion**

Our study demonstrates a significant association between COVID-19 specific search terms ‘smell loss’ and ‘taste loss’ and new weekly COVID-19 cases in Poland, particularly at one-week delay using cross correlation analysis. Subsequent time series linear regression analysis for smell loss and taste loss yielded a model with a near perfect fit, suggesting that Google Trends for these pathognomonic symptoms have a high predictive accuracy for confirmed SARS-CoV-2 infection 1-2 weeks later in Poland.
Studies investigating variations in online search preferences in the context of past disease outbreaks have repeatedly demonstrated that patterns of Google search frequencies and disease surveillance data have a striking resemblance, whereby their temporal relation could allow for prediction of future epidemiological trajectory from internet search volumes at a particular period in time. These infodemiology metrics have likewise been studied during the current COVID-19 pandemic. While some studies employed less specific symptoms such as fever, dry cough, fatigue, nasal congestion and dyspnea [9], particularly strong correlations have also been illustrated between the more pathognomonic symptoms such as anosmia and dysgeusia and future COVID-19 incidence patterns [5]. Notably, a study by Lippi et al. found that this association becomes stronger when correlating Google Trends with COVID-19 cases reported two weeks later, as opposed to cases reported in the same week [6]. Our findings support this, as both smell and taste loss produced the highest cross correlation coefficients in comparison to all non-pathognomonic symptoms, taking into account the time delay required for optimal correlation. Moreover, background seasonal patterns of diseases with similar non-specific symptoms, especially heading into the flu and cold season, may be significant confounders in search trends, underscoring the importance of relying on specific symptoms [7]. Furthermore, Lippi and Mattiuzzi stressed on the importance of considering the ‘atypical’ symptoms of anosmia and dysgeusia for case definition, particularly given the high prevalence among COVID-19 patients (49% and 48%, respectively), high positive predictive value for confirmed infection (83%), and the fact that these signs develop early after infection versus postinfection in the case of the common cold and influenza [10].

Surveillance based on Google Trends is limited by the influence of mass media communications as a possible confounder of internet user behavior. Conversely, decreasing need to research self-symptoms due to increased knowledge of anosmia and dysgeusia as COVID-19 specific symptoms may undermine the reliability of predicted infection rates.
Conclusions

In conclusion, Google Trends for anosmia and dysgeusia have a high predictive power for anticipating new COVID-19 cases 1-2 weeks ahead of official reports in Poland, serving as a useful infodemiological tool for anticipating an impending outbreak, with the potential of providing valuable buffer time to allocate necessary supplies and personnel to hospitals expecting a surge in COVID-19 patients. Upon verification by prospective research comparing model performance in different regions of Poland, public health organizations are encouraged to take advantage of this free forecasting system to anticipate and effectively manage COVID-19 outbreaks throughout Poland.

Funding: none declared.

Conflict of interest: none declared.

Human and animal rights: The study doesn’t involve human participants and/or animals, and informed consent is not required.

Ethical approval: This article does not contain any studies with human participants performed by any of the authors.

Acknowledgments: none.

Contribution statement: BMH, GL and GJ conceived the concept of the study. BMH, IS, and MHSO contributed to the design of the research. GJ and MM were involved in data collection. MHSO statistically analyzed the data. IS interpreted the data and wrote the paper. All authors edited and approved the final version of the manuscript.
References:


Figure 1. New weekly COVID-19 cases and Google Trends scores for Polish search terms “utrata węchu” (smell loss), “utrata smaku” (taste loss) (A), and non-smell and non-taste specific symptoms (B) for the period between May 18th and November 22nd, 2020, in Poland. Actual values for each week are presented in Supplementary material, Table S1. Time series regression models are shown for Google Trends of Polish search terms “utrata węchu” (smell loss) (C), and “utrata smaku” (taste loss) (D) compared with trend of reported COVID-19 cases for the period between May 18th and November 22nd, 2020, in Poland.