Master Thesis Health Economics
Erasmus University
Supervisor: Dr. F. Principe
Second assessor: Dr. C.A. Uyl-de Groot

The use of Google Trends in monitoring and predicting COVID-19 development in the Netherlands

Paul de Korte (456765)

Abstract

One of the difficulties governments face in the COVID-19 crisis is the lack of real-time data about the spread of COVID-19. Google Trends has been proposed as a reliable tool to monitor and predict infectious diseases, as Google searches reveal information about people's health. The results of fixed-effects models in this paper show that the lags of Google searches related to COVID-19 symptoms significantly correlate with COVID-19 infections, hospital admissions and deaths. Likely this means that people who recognize symptoms use Google to verify this. By also including autocorrelation, a prediction model is specified based on lasso-estimates. The prediction model shows that Google Trends can predict the development of COVID-19. Therefore, governments can use Google Trends to monitor and predict infectious diseases when accurate real-time data is missing.

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Table of content

1. Introduction

What started with a few diagnoses of a new lung disease in China, has become one of the most urgent crises of the last century. Since COVID-19 was first recognized in the Chinese city Wuhan in December 2019, the number of infections has increased rapidly. Also, the new virus has spread over all continents leading to a confirmed number of infections of 823,626 on April 1 2020, whereas the day before 72,736 new cases were reported (WHO, 2020).

COVID-19 is the name for the disease caused by the Severe Acute Respiratory Syndrome CoronaVirus 2 (SARS-CoV-2) (WHO, 2020). According to the WHO (2020), the clinical picture of COVID-19 consists of the following symptoms: *'fever, dry cough, fatigue, sputum production, shortness of breath, sore throat, headache, myalgia or arthralgia, chills, nausea or vomiting, nasal congestion, diarrhea, hemoptysis, and conjunctival congestion'.* In severe cases, COVID-19 can lead to, amongst others, viral lung infections and acute respiratory distress syndromes. The mortality rate of COVID-19 is estimated to be between 3-7% (Mehta et al., 2020; Baud et al., 2020). This novel disease is mainly fatal for men and elderly people (Surveillances, 2020). The reproduction rate of COVID-19 is on average 3.48, according to Liu et al. (2020). This number represents the average number of new infections generated by a COVID-19 patient in a naive population.

As COVID-19 is highly contagious, the disease has spread rapidly over the world. On February 23, the first case of COVID-19 was reported in the Netherlands. Since then, the Dutch government has implemented drastic measures. On March 6, inhabitants of Noord-Brabant, at that moment the province with the first confirmed cases, were advised by the RIVM to minimize social contact in cases of coughing, having a cold, or having a fever. On March 9, the Dutch government prohibited people to shake hands. Three days later, March 12, all events with more than 100 visitors were cancelled, employees in non-vital sectors were obliged to work from home if possible and universities are closed. On March 15, also schools and the catering industry are obliged to close. At that moment, 20 patients diagnosed with COVID-19 had passed away. This number increased to 5,109 on May 1 (RIVM, 2020). A maximum of Intensive Care (IC) patients was reached on April 8 (NICE, 2020), with 1,322 patients with COVID-19 being treated on IC-units. Since then, the number of IC-patients decreased, resulting in a reduction of the pressure on the Dutch healthcare.

As a result of the decreasing number of new infections in the Netherlands, the Dutch government eased the measures. On April 21, the Dutch prime minister informed the Dutch population that the primary schools were opened again on May 11. On May 6, the Dutch government further eased the measures, by presenting a roadmap of the route to open the society again. Components of this are the allowance of contact professions per May 11, the catering industry per June 1, and holiday parks per July 1. This easing of measures was on condition that the number of COVID-19 infections does not increase too fast.

To monitor this, reliable data is needed to monitor the current situation and predict the future development of COVID-19 in the Netherlands.

One of the greatest difficulties of this crisis Dutch policymakers face is the lack of data about the real number of COVID-19 infections. At the end of April only healthcare workers and vulnerable people could apply for a corona test, due to scarcity of tests. Therefore, only a small proportion of COVID-19 patients are detected in the Netherlands. This is also reflected in the observed deaths due to corona. The Dutch National Institute for Public Health and Environment (RIVM) only counts the deaths of patients diagnosed with COVID-19. However, the Statistics Netherlands (CBS) estimated for week 16 a death rate of COVID-19 which was twice as high as the deaths rate reported by the RIVM for that week (CBS, 2020). Still, the RIVM (2020) respects hospital admissions as a reliable measurement of the impact of new regulations. However, time is passing before an infected patient is hospitalized, meaning that no accurate real-time data about the coronavirus is available in the Netherlands.

A tool that is proposed to predict the progress of infections is Google Trends (Gingsberg et al., 2008). Google Trends might be a reliable epidemiologically tool as people suspecting an infection will search online for health information (e.g. symptoms) (Brunori and Resce, 2020). For this reason, Google developed a tool, called Google Flu. Between 2008-2015, this tool used real-time data to predict infections, using Google search commands. Although Google Flu showed a strong correlation between infections and Google search commands, the tool was removed after being under criticism of overpredicting epidemics. As for COVID-19, real-time data is not available in the Netherlands, Google Trends might help to both monitor and predict further development of COVID-19.

To further investigate whether Google Trends can monitor and predict the infection rate of COVID-19, the research question of this paper is:

How can Google Trends monitor and predict the development of COVID-19 in the Netherlands?

Due to the lack of real-time data, answering this research question has high social relevance. As Google Trends could spell out the effect of the taken governmental measures, this research can shed new light on the debate about effective measures. This research is also scientifically relevant, as it contributes to the complex relation between online health information and health, and it helps to understand the spread of COVID-19. Also, this paper is the first paper that assesses the use of Google Trends in predicting COVID-19 in the Netherlands.

In this paper, first the literature will be reviewed. Then the methodology used will be discussed. Afterward, the results will be presented. Lastly, a discussion of the founded results and recommendations for future research will be provided.

2. Literature review

In this literature review, first the literature about the use of Google Trends data for predicting infectious diseases will be examined. Secondly, literature focussing on patients' use of online health information will be reviewed.

As mentioned in the Introduction, with the use of Google Trends the development of infectious diseases has been predicted in the past. Ginsberg et al. (2009) were the first who investigated search engine query data to track Influenza epidemics in the US. They found a strong correlation between the number of physician visits with influenza-symptoms and search engine queries. The authors concluded that monitoring search queries might help to detect epidemics in regions where many people use the internet to search for disease-related topics. Carneiro and Mylonakis (2009) discussed the use of Google Trends in predicting regional outbreaks of diseases with high prevalence. They spelled out a theoretical framework in people searching for, in this case, influenza-related topics are patients with early symptoms of Influenza. Based on this, the authors concludes that Google Trend is a reliable real-time surveillance tool. A drawback however is that Google Trends seems to be most accurate in developed countries, as there more people use Google to search for disease-related topics. Google Trends is not only used to predict the outbreak of Influenza, but it is also used to predict the development of other diseases. Chan et al. (2011), Althouse et al. (2011) and Husnayain et al. (2019) used Google Trends to monitor the spread of Dengue fever in Bolivia, Brazil, India, Indonesia and Singapore. All these authors conclude that Google Trends is a useful tool alongside traditional surveillance of diseases. Sciascia et al. (2017) researched the use of Google Trends in the surveillance of Systemic Lupus Erythematosus on all continents. Also, these authors found positive results and recommend that Google Trends can be used as a surveillance tool for infectious diseases.

Google Trends has also been used to monitor and develop the spread of COVID-19. Hu et al. (2020) studied Google search queries in six English-speaking countries. They compared the number of daily queries related to COVID-19 with the number of people infected with COVID-19. The authors found a small positive correlation between both variables. The authors focussed on measuring the public awareness of COVID-19 and not on monitoring and predicting COVID-19, so the authors did not further investigate the measured correlation. Husnayain et al. (2020) researched whether Google Trend can be used to monitor COVID-19 in Taiwan. They found a high to moderate correlation between the number of positive diagnoses and search queries related to COVID-19. They concluded that Google Trends can be a useful tool to define risk communication strategies. Li et al. (2020) in retrospect, tried to predict the COVID- outbreak based on search engine queries in China. Using internet searches and social media data, Li et al. could predict the development of COVID-19 10-14 days in advance. The keywords these authors used were 'coronavirus' and 'pneumonia'. Brunori & Resce (2020) found a lag of 6-10 days

between Google searches regarding COVID-19 symptoms and COVID-19 deaths in Italy. Similar results were obtained by Farzanegan et al. (2020) for Iran.

The relation between Google search queries and people's health also sheds light on the role of online health information in today's society. The literature specifies two opposed effects of online health information seeking on general health. The first one is that access to online health information improves the knowledge of patients, whereas they can discuss their treatment plan more actively with physicians. The literature indicates that health information is complementary to physician visits, not a substitution. The second effect is that online health information is difficult to understand, especially for elder people and low-educated.

Suziedelyte (2012) studied the relation between online health information seeking and the number of visits to a health professional in the US. For this, Suziedelyte used an Instrumental Variable (IV) approach where US state telecom regulations were used as an instrument for a dummy-variable which indicated whether or not an individual used the internet recently to search for health information. The state regulations that were used as an instrument influenced the quality of internet access randomly per state. Suziedelyte found that individuals who used the internet to search for health information recently demanded more health care. Therefore, she concludes that online health information is complementary to formal health care. Suziedelyte highlights a potential pathway how online health information seeking could influence health. She shows that searching for online health information influences an individual's number of visits to a health professional. This perspective is further evaluated by Tan and Goonawardene (2017) in a literature study of 18 published articles. Their focus is on internet health information seeking and the patient-physician relationship. In line with Suziedelyte (2012), Tan and Goonawardene concluded that online health information seeking strengthens the relationship between the patient and the physician. This conclusion is shared by McMullen (2006). She adds to this that patients who searched on the internet for health information participated more actively in the discussion about a treatment plan.

On the other hand, the literature indicates that health information on the internet is difficult to understand. Benigeri and Pluye (2003) noted that access to digital health information favors high-educated. Also, elderly people do not have good access to online health information, while they have many health problems. Tonsaker et al. (2014) concluded that online health information is still unclear for many individuals. Therefore, online health information might be less valuable to elderly and lower educated people.

This research will mainly focus on the first pathway, as it explores the relation between the use of online health information and the number of COVID-19 indications. Due to a lack of data, it is not possible to explore heterogeneity for elderly and low-educated people.

3. Methodology

In the literature different key explanatory variables are used. Farzanegan et al. (2020) use the number of search queries containing either 'Corona Symptoms', 'Masks', 'Disinfection', or 'Corona'. For both the lags of the number of search queries containing 'Corona Symptoms' and 'Disinfection' the authors find positive correlations with the number of confirmed cases of COVID-19. Brunori & Resce (2020) use the number of search queries for some of the most commonly reported symptoms: 'Fever', 'dry cough', 'sore throat', 'loss of sense of smell' and 'loss of sense of taste'. As the authors find a certain degree of heterogeneity between the number of search queries on Google for these symptoms, they normalize the results and create a sum-variable with a range of 0-100. The conclusion of Brunori & Resce (2020) is that Google Trends data has a strong ability to predict the number of COVID-19 deaths in Italy. Hu et al. (2020) use the following search queries to identify the public awareness of COVID-19: 2019-nCoV', 'SARS-CoV-2', 'novel coronavirus', 'new coronavirus', 'COVID-19' and 'Corona Virus Disease 2019'. Lastly, Li et al. (2020) assessed the correlation between the number of search queries containing 'Coronavirus' and 'Pneumonia' and the number of confirmed cases of COVID-19. They found the strongest correlation for the term 'Coronavirus'.

To conclude, the literature specifies three possibilities to use Google Trends to monitor and predict COVID-19. The first possibility is to use the number of search queries on Google for COVID-19 itself (Corona, COVID-19, SARS-CoV-2, etc.). Secondly, it is possible to use search queries about corona symptoms (Corona Symptoms, fever, dry cough, etc.). Lastly, it is also possible to use search queries related to the fight against COVID-19 (masks, disinfection, etc.). All three possibilities will be further explored in this paper.

Several variables can be used as outcome variables. First of all, the number of confirmed cases of COVID-19 can be used. This is done in most other studies. However, as aforementioned, in the Netherlands, the test capacity has been too small to test everyone. A second option is to use the number of hospital admissions related to COVID-19. The RIVM regards this as a reliable indication of the total number of COVID-19 patients. A third option is the number of confirmed deaths due to COVID-19. However, as aforementioned, due to the lack of test capacity, the number of confirmed deaths is likely to be an underestimation of the total number of deaths due to COVID-19. In this paper, all possibilities will be explored and evaluated. As no regional data about IC admissions is available, this will not be used as an outcome variable.

To investigate the correlation between search queries on Google Trends, regional fixed-effects regression models will be measured, as done by e.g. Farzanegan et al. (2020). The specification of these models will be further optimized by using lasso estimates. Lastly, also regional fixed effects models are estimated whereas the lags of the dependent variable are included to predict the number of COVID-19 cases, hospital admissions and deaths more precisely (See e.g. Morsy et al. (2018)). Fixed-effects models

are used as these models focus on within-regional differences and therefore take unobserved time-invariant variables into account.

The model specification is based on least absolute shrinkage and selection operator (lasso) estimations. Lasso adds a penalty to each coefficient which is equal to the magnitude of the coefficient, while it minimizes the mean squared prediction error. The weight of this parameter is determined by the parameter λ. Due to the penalty term, a more parsimonious model can be obtained, whereas the best predictors are included in the model (Tibshirani, 1996). For each model cross-validation, adaptive and plug-in lasso coefficients are estimated for out-of-sample predictions. The model with the lowest MSE and highest R^2 is chosen.

As panel data will be used, it is possible to control for time-invariant unobserved variables. A method to do is, is the fixed effects regression. This model has the following functional form:

$$Y_{it} = \beta_0 + \sum_{k=0}^{n} \beta_{it-k}\ General_{it-k} + \sum_{k=0}^{n} \theta_{it-k}\, Symptoms_{it-k} + \sum_{k=0}^{n} \Phi_{it-k}\ Measures_{it-k} + \varepsilon_i$$

Here Y_{it} is the number of confirmed COVID-19 cases, deaths, or hospital admissions, as all three options will be assessed. $General_{\text{it-k}}$ is the number of Google searches for COVID-19 in general in region *i* on day *t*, with *k* as the lead. $Symptoms_{\text{it-k}}$ is the number of Google searches for COVID-19 symptoms in region *i* on day *t*, with *k* as the lead. $Measures_{\text{it-k}}$ is the number of Google searches for COVID-19 measures in region *i* on day *t*, with *k* as the lead. Forty lags will be included, to account for the time between the incubation of the virus and deaths (see Verity et al., 2020). To further assess which variables are the best predictors, lasso estimations are used. As some newspapers reported that in the northern provinces more people were tested. Therefore, the described models will be estimated separately for the northern provinces.

Next to this, a fixed-effects model is estimated for which autocorrelation is included. The functional form is then:

$$Y_{it} = \beta_0 + \sum_{k=0}^{n} \beta_{it-k}\ General_{it-k} + \sum_{k=0}^{n} \theta_{it-k}\, Symptoms_{it-k} + \sum_{k=0}^{n} \Phi_{it-k}\ Measures_{it-k} + \sum_{j=0}^{n} \mu_{it-j}\ Y_{it-j} + \varepsilon_i$$

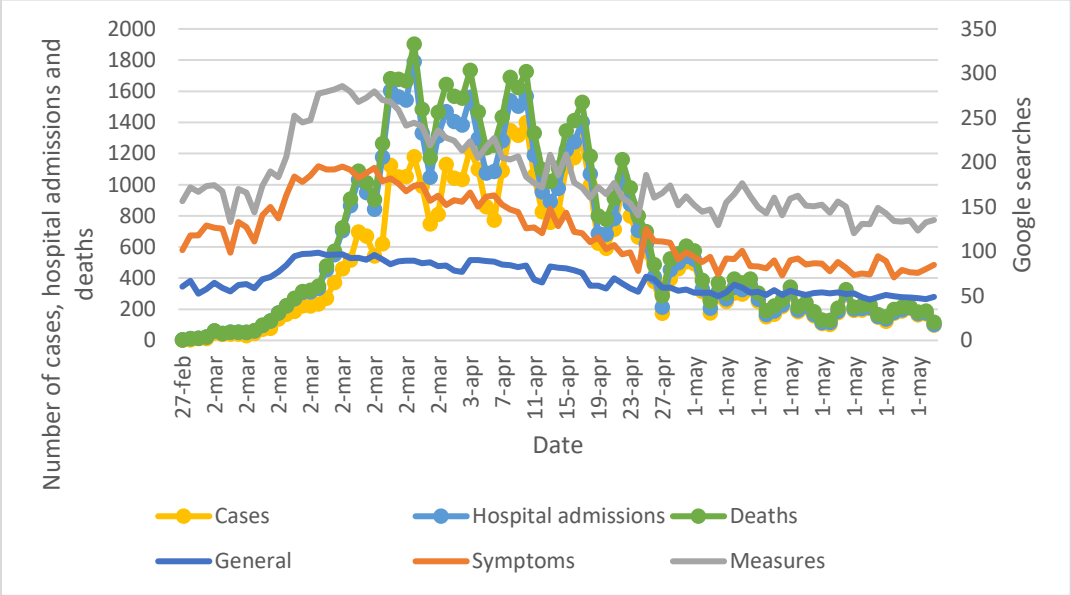
The same variables are used as before, whereas also lags of the dependent variable are included, with *j* as the lead. Again, the number of lags included will be based on lasso estimations.

4. Data

Data on COVID-19 cases, deaths and hospital admissions are reported daily by the RIVM. By clustering communities, data on province level was obtained. On some days, the RIVM reported a negative number of new infections, hospital admissions and deaths. This was done to correct the reported number of infections, hospital admissions and deaths of the day before. To correct this, both negative values and the observations of the day before the negative values were reported as missing.

Google Trends data is retrieved from Google Trends. The data derived from Google Trends is normalized and clustered. Three variables are constructed, which represent the number of Google search queries per day related to (1) COVID-19 in general (*General*), (2) COVID-19 symptoms (*Symptoms*) and (3) COVID-19 measures (*Measures*) in the period of 1 December 2019 until 31 May 2020. The variable which indicates the number of Google search queries related to COVID-19, in general, is based on the Google search queries for the following terms: "COVID-19", "Corona" and "Coronavirus". As already mentioned, these search queries are also used in the literature. The variable which indicates the number of Google search queries related to COVID-19 symptoms is based on the Google search queries for the most frequent symptoms of COVID-19 according to the WHO (2020). These are: "Corona symptomen" (Dutch for "Corona symptoms"), "Koorts" ("fever"), "Droge hoest" ("dry cough"), "Slijm" ("sputum") and "Benauwdheid" ("stuffiness"). Lastly, the variable indicating the search queries related to COVID-19 measures is partly based on the literature and partly based on the most important measures taken by the Dutch government. These are: "Corona maatregelen" (Dutch for "Corona measures"), "Lock down", "Social distancing", "Mondkapje" ("face mask") and 'Desinfectie" ("disinfection"). The scores for the three variables are between 0 and 100, whereas a higher score indicates more public interest. The range of the score is the same as the initial Google Trends data. Figure 1 shows the course of the data.

Figure 1: Number of cases, hospital admissions and deaths per day together with Google searches about COVID-19 in general, COVID-19 symptoms and COVID-19 measures in the Netherlands



5. Results

To investigate the relation between the number of COVID-19 cases, hospital admissions and Google search queries related to COVID-19, three regional fixed effects models are estimated. The results are shown in Appendix A, Table 4-6.

Table 4 shows the results of the fixed effects analysis with the number of COVID-19 as the dependent variable and the Google search queries as the independent variables. For *General* and *Measures*, lags are only incidental significantly correlated with the number of COVID-19. For *Symptoms*, almost all lags between the 10th and 32nd lag are significantly and positively correlated with the number of COVID-19 cases. The magnitude of the coefficients for the significant lags is between 0.14 and 0.34, indicating that on average an increase of the Google Trends score for COVID-19 with 1 goes hand in hand with an increase of 0.10-0.34 COVID-19 cases, 10 to 32 days later.

Table 5 shows the estimations for the regional fixed effects model with COVID-19 hospital admissions as the dependent variable and the Google Trends data as independent variables. For *General* and *Measures*, only some lags are significantly correlated with the number of COVID-19. For *Symptoms*, almost all lags between the 19th and 32nd lag are significantly and positively correlated with the number of COVID-19 cases. The magnitude of the coefficients for the significant lags is between 0.08 and 0.14, indicating that on average an increase of the Google Trends score for COVID-19 with 1 is associated with an increase of 0.08-0.15 hospital admissions due to COVID-19, 19 to 32 days later, all ells being equal.

Table 6 shows the estimations for the regional fixed effects model with COVID-19 deaths as dependent variable *General*, *Measures* and *Symptoms*, with each forty lags, as independent variables. For *General* and *Measures*, lags are only incidental significantly correlated with the number of COVID-19. For *Symptoms*, almost all lags between the 23rd and 33rd lag are significantly and positively correlated with the number of COVID-19 cases. The magnitude of the coefficients for the significant lags is between 0.08 and 0.14, indicating that on average an increase of the Google Trends score for COVID-19 with 1 is correlated with an increase of 0.04-0.08 deaths due to COVID-19, 23 to 32 days later, ceteris paribus.

In all three estimations, a series of lags of *Symptoms* is significantly correlated with the independent variable. For the number of COVID-19 infections, the most and earliest lags are significantly correlated, whereas the magnitude of the coefficients is the highest. Compared to the estimation in Table **3**, the second estimation shows more and earlier significant coefficients of the lags of *Symptoms*, and also the magnitude of these coefficients is larger. Further analysis shows that these results were mainly driven by the Google search queries for sputum. Furthermore, when the same models were estimated for the three northern provinces (Groningen, Friesland and Drenthe) no series of significant estimates were found.

To further analyse which lags of *General*, *Measures* and *Symptoms* can predict the number of COVID-19 infections, hospital admissions and deaths in the Netherlands, the least absolute shrinkage and selection operator (lasso) is used.

As the maximum incubation time is around 14 days, an adaptive lasso is estimated for a predicting model with the number of COVID-19 as the dependent variable and *General*, *Measures* and *Symptoms* with 14 lags each, as independent variables. The model specification based on the covariates chosen by the adaptive lasso is as follows:

Table 1: Regional fixed-effects model with the number of COVID-19 cases as the dependent variable and Google searches as independent variables, with maximally 14 lags.

Cases	**Coefficient (standard deviation)**
General (T-6)	0.261 (0.102)**
General (T-14)	0.548 (0.093)***
Symptoms	-0.332 (0.088)***
Symptoms (T-1)	-0.327 (0.090)***
Symptoms (T-10)	0.223 (0.095)**
Symptoms (T-11)	0.333 (0.095)***
Symptoms (T-12)	0.374 (0.095)***
Symptoms (T-13)	0.350 (0.095)***
Symptoms (T-14)	0.435 (0.094)***

Measures (T-2)	0.069 (0.086)
Measures (T-9)	-0.085 (0.101)
Measures (T-10)	0.020 (0.103)
Measures (T-14)	0.021 (0.094)
Constant	-69.279 (7.147)***

Table 1: #Observations: 1,130. Model specification is based on adaptive lasso-estimates. The standard errors are reported in parentheses. * p < 0.1; ** p < 0.05; *** p < 0.01.

The adaptive lasso estimates included lags of *General* and *Measures* incidentally, but for *Symptoms* a series of lags are included. Lag 10 until lag 14 of *Symptoms* are included, whereas the fixed-effects model shows positive significant estimates for these lags. Remembering the results of the regional fixed-effects model where even the 32nd lag of *Symptoms* was significant, also lasso coefficients were measured for the same model but now with 32 lags of the independent variable. The outcome of the fixed effects model based on the covariates selected by the adaptive lasso can be found in Appendix B**,** Table 7. For the variable *General*, lasso estimates selected, next to some incidental lags, a series of lags around the 10th to 25th lag. For Measures, most of the first twenty lags are included. However, the fixed-effects model provides only incidentally significant coefficients for *General* and *Measures*. For *Symptoms*, almost all lags, from lag 10 onwards, are included. These coefficients are mostly positive and significant in the fixed-effects analysis.

To estimate which lags are most important in predicting the number of hospital admissions due to COVID-19 lasso is used to select out of forty lags of the three independent variables, the most important lags. The results of the regional fixed-effects model with the coefficients included as suggested by the lasso estimates are as follows:

Table 2: Regional fixed-effects model with the number of hospital admissions due to COVID-19 as the dependent variable and Google searches as independent variables, with maximally 40 lags.

Hospital admissions	**Coefficients (standard deviation)**
General	-0.199 (0.070)***
General (T-2)	-0.033 (0.069)
General (T-5)	0.018 (0.070)
General (T-6)	-0.093 (0.070)
General (T-12)	-0.040 (0.069)
General (T-13)	-0.125 (0.069)*
General (T-14)	-0.037 (0.070)
General (T-15)	-0.079 (0.068)
General (T-19)	-0.085 (0.066)
General (T-20)	-0.077 (0.067)

General (T-21)	-0.105 (0.068)
General (T-22)	-0.103 (0.069)
General (T-23)	-0.094 (0.070)
General (T-24)	-0.167 (0.069)**
General (T-25)	-0.061 (0.070)
General (T-26)	-0.151 (0.066)**
General (T-32)	0.119 (0.067)*
General (T-33)	0.112 (0.072)
General (T-34)	0.050 (0.072)
General (T-35)	0.046 (0.070)
Symptoms (T-1)	0.057 (0.052)
Symptoms (T-2)	0.117 (0.052)**
Symptoms (T-3)	0.079 (0.052)
Symptoms (T-5)	0.138 (0.051)***
Symptoms (T-7)	0.027 (0.051)
Symptoms (T-8)	0.069 (0.050)
Symptoms (T-9)	0.103 (0.051)**
Symptoms (T-10)	0.120 (0.051)**
Symptoms (T-11)	0.097 (0.051)*
Symptoms (T-12)	0.134 (0.050)***
Symptoms (T-13)	0.124 (0.050)**
Symptoms (T-16)	0.011 (0.050)
Symptoms (T-17)	0.032 (0.050)
Symptoms (T-19)	0.106 (0.049)**
Symptoms (T-20)	0.101 (0.049)**
Symptoms (T-21)	0.150 (0.049)***
Symptoms (T-22)	0.086 (0.048)*
Symptoms (T-23)	0.125 (0.048)***
Symptoms (T-24)	0.112 (0.049)**
Symptoms (T-25)	0.082 (0.049)*
Symptoms (T-26)	0.091 (0.049)*
Symptoms (T-27)	0.096 (0.049)*
Symptoms (T-29)	0.130 (0.049)***
Symptoms (T-30)	0.068 (0.049)
Symptoms (T-31)	0.092 (0.048)**

Symptoms (T-32)	0.102 (0.048)**
Symptoms (T-33)	0.065 (0.049)
Symptoms (T-36)	-0.077 (0.048)
Symptoms (T-37)	-0.133 (0.048)***
Measures	0.014 (0.056)
Measures (T-1)	-0.013 (0.055)
Measures (T-2)	0.122 (0.056)**
Measures (T-3)	0.081 (0.055)
Measures (T-4)	0.022 (0.055)
Measures (T-5)	0.048 (0.055)
Measures (T-6)	0.103 (0.055)*
Measures (T-7)	0.031 (0.055)
Measures (T-8)	0.085 (0.055)
Measures (T-9)	0.053 (0.055)
Measures (T-10)	0.091 (0.055)*
Measures (T-11)	0.063 (0.055)
Measures (T-12)	0.086 (0.055)
Measures (T-13)	0.190 (0.056)***
Measures (T-14)	0.024 (0.056)
Measures (T-15)	0.092 (0.056)*
Measures (T-16)	0.039 (0.055)
Measures (T-17)	0.056 (0.055)
Measures (T-18)	-0.038 (0.055)
Measures (T-27)	-0.111 (0.055)**
Measures (T-30)	-0.092 (0.057)
Measures (T-31)	-0.066 (0.058)
Measures (T-35)	0.064 (0.058)
Measures (T-37)	0.050 (0.059)
Measures (T-38)	0.091 (0.059)
Measures (T-40)	0.062 (0.056)
Constant	-108.229 (14.255)***

Table 2: #Observations: 932. Model specification is based on cross-validation lasso-estimates. The standard errors are reported in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Cross-validation lasso estimates selected for *General* a series of lags between the 12th lag and the 35th lag while for *Measures* mostly early lags are included. However, most of these lags are not significant in the regional fixed-effect model. For *Symptoms*, almost all lags between lag 19 and lag 32 are included.

These coefficients are mostly significant and positive. As this model is based on cross-validation lasso estimates, the model is less parsimonious compared to a model based on adaptive or plugin lasso estimates.

For exploring which lags are best at predicting the number of COVID-19 deaths, the same steps are repeated. The results of the fixed effects model with the parameters included as suggested by the lasso estimates can be seen in Table 3.

Table 3: Regional fixed-effects model with the number of COVID-19 deaths as the dependent variable and Google searches as independent variables, with maximally 40 lags.

Deaths	**Coefficient (standard deviation)**
General	-0.015 (0.030)
General (T-1)	-0.013 (0.030)
General (T-2)	-0.018 (0.029)
General (T-4)	-0.010 (0.028)
General (T-6)	-0.020 (0.029)
General (T-7)	0.018 (0.029)
General (T-9)	0.014 (0.028)
General (T-14)	-0.028 (0.028)
General (T-15)	-0.044 (0.028)
General (T-17)	-0.038 (0.028)
General (T-19)	-0.062 (0.028)**
General (T-20)	0.013 (0.028)
General (T-21)	-0.056 (0.028)**
General (T-22)	-0.026 (0.029)
General (T-23)	-0.012 (0.029)
General (T-24)	-0.031 (0.028)
General (T-26)	-0.019 (0.028)
General (T-27)	-0.069 (0.028)**
General (T-34)	-0.008 (0.027)
General (T-37)	0.004 (0.028)
General (T-39)	-0.048 (0.028)*
Symptoms	-0.066 (0.021)***
Symptoms (T-2)	0.007 (0.021)
Symptoms (T-6)	-0.019 (0.021)
Symptoms (T-10)	0.072 (0.021)***

Symptoms (T-11)	-0.002 (0.021)
Symptoms (T-12)	0.038 (0.020)*
Symptoms (T-13)	0.021 (0.020)
Symptoms (T-15)	-0.009 (0.020)
Symptoms (T-16)	0.011 (0.021)
Symptoms (T-17)	0.030 (0.020)
Symptoms (T-18)	0.027 (0.021)
Symptoms (T-19)	0.041 (0.020)
Symptoms (T-20)	0.027 (0.020)
Symptoms (T-21)	0.023 (0.020)
Symptoms (T-23)	0.050 (0.020)**
Symptoms (T-24)	0.055 (0.020)***
Symptoms (T-25)	0.040 (0.020)**
Symptoms (T-26)	0.043 (0.020)**
Symptoms (T-27)	0.061 (0.020)***
Symptoms (T-28)	0.044 (0.020)***
Symptoms (T-29)	0.031 (0.020)
Symptoms (T-30)	0.069 (0.020)***
Symptoms (T-31)	0.049 (0.020)**
Symptoms (T-32)	0.044 (0.020)**
Symptoms (T-33)	0.037 (0.020)*
Symptoms (T-34)	0.017 (0.020)
Symptoms (T-35)	0.027 (0.020)
Symptoms (T-36)	0.013 (0.020)
Symptoms (T-37)	0.019 (0.020)
Symptoms (T-38)	0.052 (0.020)**
Symptoms (T-39)	-0.010 (0.020)
Symptoms (T-40)	0.014 (0.020)
Constant	-30.988 (6.568)***
Measures	0.024 (0.023)
Measures (T-1)	0.042 (0.023)*
Measures (T-3)	0.007 (0.023)
Measures (T-4)	-0.010 (0.023)
Measures (T-5)	-0.003 (0.023)
Measures (T-6)	0.034 (0.023)

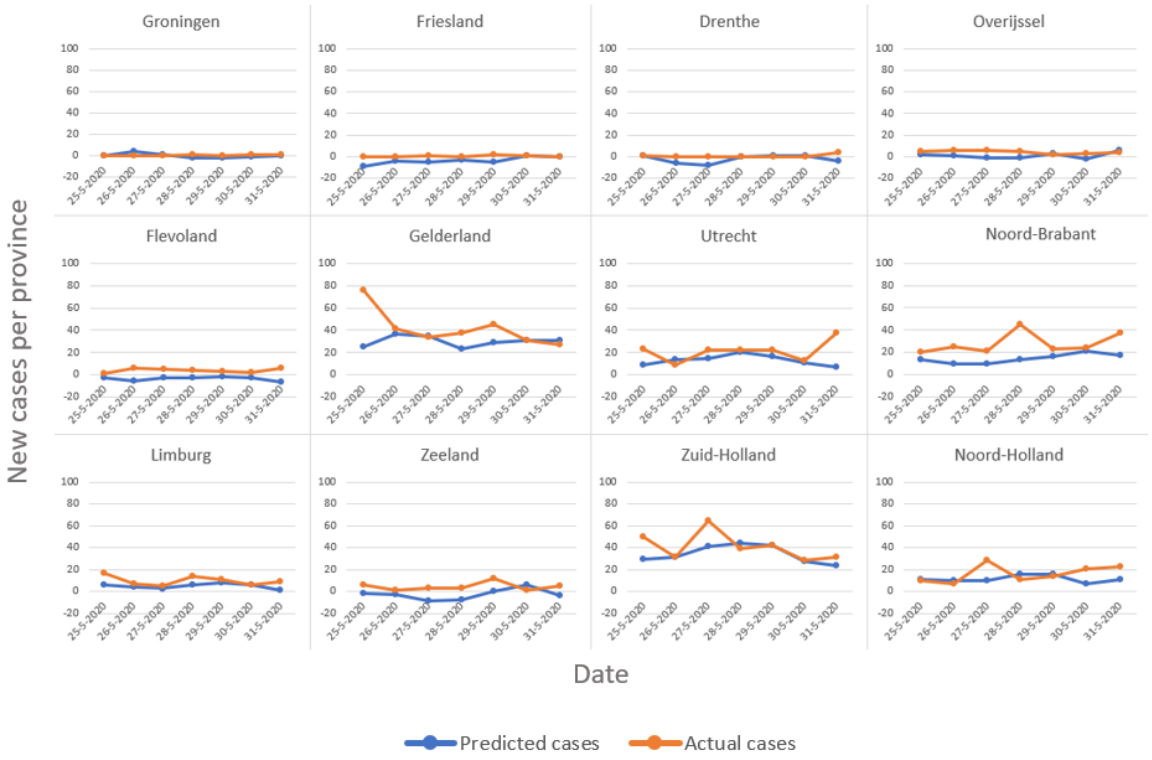
Measures (T-7)	0.024 (0.023)
Measures (T-8)	0.016 (0.023)
Measures (T-9)	0.028 (0.023)
Measures (T-10)	0.000 (0.023)
Measures (T-11)	0.011 (0.023)
Measures (T-12)	0.001 (0.023)
Measures (T-13)	0.019 (0.023)
Measures (T-14)	0.021 (0.023)
Measures (T-15)	0.064 (0.023)***
Measures (T-16)	0.048 (0.023)**
Measures (T-17)	0.008 (0.023)
Measures (T-18)	-0.002 (0.023)
Measures (T-20)	-0.010 (0.023)
Measures (T-21)	0.006 (0.023)
Measures (T-22)	0.030 (0.023)
Measures (T-23)	0.018 (0.023)
Measures (T-24)	-0.019 (0.023)
Measures (T-27)	-0.047 (0.023)
Measures (T-30)	-0.025 (0.023)
Measures (T-33)	0.014 (0.023)
Measures (T-35)	-0.024 (0.024)
Measures (T-37)	0.026 (0.025)
Measures (T-39)	0.000 (0.025)
Measures (T-40)	0.037 (0.024)

Table 3: #Observations: 946. Model specification is based on cross-validation lasso-estimates. The standard errors are reported in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Compared to Table 2 the series of lags included are not very different for each of the independent variables. For each of the three variables, some later lags are added, compared to Table 2. The significant coefficients are mostly the 23rd until the 32nd lag of *Symptoms*. These significant coefficients are all positive.

To predict the number of COVID-19 cases per province accurately also lags of the dependent variable are added to the regional fixed effects model. The specification of this model is based on plugin lasso estimates, as these estimates led to the lowest MSE for out-of-sample predictions. The pseudo-out-of-sample prediction for the last week of May differed on average 6.3 cases with the real number of COVID-19 cases. In Graph 2 the predicted and real number of COVID-19 cases per day per province is plotted.

Graph 2: Actual number of new infections per day per province compared to predicted number of infections per day per province.



Graph 2: #Observations: 253. Prediction based on plugin lasso-estimates. R^2: 0.818; MSE: 710.37. Average number of cases in last week of May: 13.8.

The same steps were repeated to predict the number of COVID-19 hospital admissions and deaths per province. The pseudo-out-of-sample prediction for the last week of May differed on average 0.33 admissions with the real number of COVID-19 hospital admissions, while the number of deaths differed with 0.48 from the real observations. Plots of the predicted and real number of COVID-19 hospital admissions and deaths per day per province are shown in Appendix C, Graph 3 and 4.

6. Conclusion

The research question of this paper was:

How can Google Trends monitor and predict the future development of COVID-19 in the Netherlands?

To answer this research question, the relationship between Google searches and COVID-19 infections, hospital admissions and deaths in the Netherlands is assessed. For this regional fixed effects models are used. The results showed that Google search queries related to COVID-19 symptoms showed a more significant correlation with COVID-19 cases, hospital admissions or deaths, than the other Google searches. For the number of COVID-19 cases, the coefficients of Google searches after COVID-19 symptoms 10 to 32 days before were mostly positive and significant. Also, the lasso-estimates included this series of lags. Google searches after COVID-19 symptoms 19 to 32 days before were mostly significantly correlated with COVID-19 hospital admissions. Lasso estimates included these series of lags for *Symptoms* and *General*, while for *Measures* earlier lags were included. However, the coefficients of the lags of *Measures* were not significant. For the number of COVID-19 deaths, Google searches after COVID-19 23 to 32 days before were mostly positive and significantly correlated. Based on these findings, the theoretical framework suggests that after someone uses Google to find more information about the symptoms of COVID-19 he or someone close to him has is tested on COVID-19 10 to 32 days later. Also, the results suggest that the moment a patient is tested (10-32 days after Googling) is close to the moment of hospital admission (19-32 days after Googling) and the moment of death (23-32 days after Googling). Compared to other literature, the founded gap between searching for online health information and being tested, hospitalized, or dying is large. Brunori and Resce (2020) found a delay between Google searches and COVID-19 deaths of 10 days in Italy while Husnayain et al. (2020) found a lag of 1 to 3 days between Google searches and COVID-19 infections. The results of Li et al. (2020) were more in line with the results in this paper. Li et al. found a lag of 10 to 14 days between online health information searches in China and the number of infections in China. A likely explanation for the founded results is that in the Netherlands not everyone could get a COVID-19 test before June 1. Unless people had a so-called crucial job they were only tested in hospital. This also explains why the results suggest that the time between being tested, being hospitalized and dying is so short. Another argument for this conclusion is that in the northern provinces, where more people could get a test, no such relationship was found.

The prediction models performed well, especially for the number of hospital admissions and number of deaths. Likely this is the case as not all patients could get a COVID-19 test.

For policy makers, this paper shows that Google Trend can be a helpful tool to monitor and predict infectious diseases if accurate information is missing. However, a shortcoming is that by using Google Trends as information source, the users can not be tracked. This makes it difficult to implement a so-called track-and-trace policy in which people who had close contact with COVID-19 patients are asked

to stay in quarantine. However, governments can still use Google Trends as a tool to plan both easing and reinforcement of measures.

This paper also has some shortcomings. First of all, the obtained results might suffer from time-varying omitted variables which are both correlated with the dependent variable and independent variables. An example of such a variable might be the number of press articles about COVID-19. Also, the results might reflect autocorrelation, as COVID-19 is an infectious disease. If someone has symptoms of COVID-19 and uses Google to verify this, he might infect others. It is therefore not possible to interpret the results as a specific time between getting symptoms and being tested for the same patient. As not all patients could be tested before June 1, but only a subsample, there might be selection bias. Lastly, the data contained observations of a negative number of new infections, hospital admissions and deaths. Although these observations were corrected, the data might still contain more mismeasurements.

Future research should research the relation between Google search queries and the development of COVID-19 for the period after June 1, as than COVID-19 tests were available for all inhabitants of the Netherlands. Secondly, future research is needed in which individuals health is linked to the search queries of these individuals. This can further spell-out the relation between the number of COVID-19 infections and online search behaviour.

All in all, this paper highlights the importance of online health information. Patients seem to use Google to verify if they have COVID-19. The government should, therefore, focus on providing clear online health information to all people. Secondly, the results of this paper show that Google Trends can be a useful tool to monitor and predict infectious diseases as COVID-19 when accurate real-time data is missing. Lastly, this paper shows the urge for fast testing to stop the spread of COVID-19 as was done in the northern provinces of the Netherlands.

Appendix A

Table 4. Regional fixed-effects model with the number of COVID-19 cases as the dependent variable and Google searches as independent variables.

Cases	Coefficient (standard deviation)
General	-0.049 (0.126)
Symptoms	0.032 (0.088)
Measures	0.088 (0.096)
General (T-1)	-0.132 (0.128)
General (T-2)	0.123 (0.128)
General (T-3)	-0.049 (0.127)
General (T-4)	0.075 (0.126)
General (T-5)	0.052 (0.126)
General (T-6)	-0.027 (0.126)
General (T-7)	-0.100 (0.125)
General (T-8)	-0.018 (0.126)
General (T-9)	-0.183 (0.126)
General (T-10)	-0.028 (0.128)
General (T-11)	-0.119 (0.128)
General (T-12)	-0.183 (0.129)
General (T-13)	-0.019 (0.129)
General (T-14)	0.013 (0.129)
General (T-15)	-0.160 (0.130)
General (T-16)	-0.163 (0.132)
General (T-17)	-0.302 (0.132)**
General (T-18)	-0.133 (0.132)
General (T-19)	-0.240 (0.134)*
General (T-20)	-0.268 (0.134)**
General (T-21)	-0.180 (0.134)
General (T-22)	-0.272 (0.136)**
General (T-23)	-0.189 (0.137)
General (T-24)	-0.145 (0.138)
General (T-25)	-0.272 (0.138)**
General (T-26)	-0.161 (0.138)
General (T-27)	-0.098 (0.140)
General (T-28)	0.049 (0.141)
General (T-29)	0.011 (0.141)
General (T-30)	0.056 (0.142)
General (T-31)	0.057 (0.141)

General (T-32)	-0.055 (0.141)
General (T-33)	-0.155 (0.141)
General (T-34)	-0.021 (0.143)
General (T-35)	0.021 (0.144)
General (T-36)	-0.151 (0.144)
General (T-37)	-0.018 (0.148)
General (T-38)	0.067 (0.149)
General (T-39)	-0.135 (0.150)
General (T-40)	0.115 (0.140)
Symptoms (T-1)	0.006 (0.088)
Symptoms (T-2)	0.017 (0.088)
Symptoms (T-3)	0.053 (0.088)
Symptoms (T-4)	0.041 (0.087)
Symptoms (T-5)	0.153 (0.088)*
Symptoms (T-6)	0.020 (0.087)
Symptoms (T-7)	0.063 (0.087)
Symptoms (T-8)	0.082 (0.087)
Symptoms (T-9)	0.099 (0.087)
Symptoms (T-10)	0.205 (0.087)**
Symptoms (T-11)	0.213 (0.087)**
Symptoms (T-12)	0.201 (0.087)**
Symptoms (T-13)	0.151 (0.087)*
Symptoms (T-14)	0.173 (0.087)**
Symptoms (T-15)	0.104 (0.087)
Symptoms (T-16)	0.178 (0.087)**
Symptoms (T-17)	0.294 (0.087)***
Symptoms (T-18)	0.277 (0.087)***
Symptoms (T-19)	0.316 (0.087)***
Symptoms (T-20)	0.271 (0.087)***
Symptoms (T-21)	0.318 (0.087)***
Symptoms (T-22)	0.337 (0.087)***
Symptoms (T-23)	0.205 (0.087)**
Symptoms (T-24)	0.320 (0.088)***
Symptoms (T-25)	0.279 (0.087)***
Symptoms (T-26)	0.188 (0.087)**
Symptoms (T-27)	0.276 (0.087)***
Symptoms (T-28)	0.189 (0.087)**
Symptoms (T-29)	0.215 (0.087)**
Symptoms (T-30)	0.195 (0.088)**

Symptoms (T-31)	0.195 (0.088)**
Symptoms (T-32)	0.205 (0.088)**
Symptoms (T-33)	0.078 (0.088)
Symptoms (T-34)	0.016 (0.087)
Symptoms (T-35)	0.105 (0.088)
Symptoms (T-36)	0.027 (0.088)
Symptoms (T-37)	-0.007 (0.088)
Symptoms (T-38)	0.034 (0.088)
Symptoms (T-39)	0.131 (0.088)
Symptoms (T-40)	0.148 (0.088)*
Measures (T-1)	0.283 (0.098)***
Measures (T-2)	0.299 (0.100)***
Measures (T-3)	0.167 (0.101)*
Measures (T-4)	0.149 (0.101)
Measures (T-5)	0.077 (0.102)
Measures (T-6)	-0.030 (0.102)
Measures (T-7)	0.176 (0.103)*
Measures (T-8)	0.165 (0.104)
Measures (T-9)	0.158 (0.104)
Measures (T-10)	0.338 (0.104)***
Measures (T-11)	0.176 (0.105)**
Measures (T-12)	0.044 (0.106)
Measures (T-13)	0.110 (0.106)
Measures (T-14)	0.122 (0.106)
Measures (T-15)	0.058 (0.107)
Measures (T-16)	0.159 (0.107)
Measures (T-17)	0.150 (0.108)
Measures (T-18)	0.066 (0.108)
Measures (T-19)	-0.074 (0.108)
Measures (T-20)	0.198 (0.108)*
Measures (T-21)	0.128 (0.109)
Measures (T-22)	0.049 (0.109)
Measures (T-23)	0.081 (0.109)
Measures (T-24)	0.140 (0.109)
Measures (T-25)	0.072 (0.110)
Measures (T-26)	-0.039 (0.112)
Measures (T-27)	0.130 (0.112)
Measures (T-28)	-0.016 (0.114)
Measures (T-29)	0.088 (0.114)

Measures (T-30)	0.043 (0.115)
Measures (T-31)	0.218 (0.116)*
Measures (T-32)	0.069 (0.116)
Measures (T-33)	0.024 (0.116)
Measures (T-34)	0.070 (0.117)
Measures (T-35)	-0.041 (0.118)
Measures (T-36)	0.028 (0.119)
Measures (T-37)	0.075 (0.120)
Measures (T-38)	0.157 (0.119)
Measures (T-39)	0.140 (0.118)
Measures (T-40)	0.128 (0.115)
Constant	-339.081 (22.060)***

Table 4: #Observations: 1,130. The standard errors are reported in parentheses. * p < 0.1; ** p < 0.05; *** p < 0.01.

Table 5. Regional fixed-effects model with the number of hospital admissions of COVID-19 patients as dependent variable and Google searches as independent variables.

Hospital admissions	**Coefficient (standard deviation)**
General	-0.180 (0.076)**
Symptoms	-0.047 (0.053)
Measures	0.005 (0.059)
General (T-1)	-0.046 (0.078)
General (T-2)	-0,033 (0,077)
General (T-3)	-0,088 (0,076)
General (T-4)	-0,026 (0,076)
General (T-5)	0,014 (0,076)
General (T-6)	-0,107 (0,076)
General (T-7)	-0,025 (0,076)
General (T-8)	0,024 (0,076)
General (T-9)	-0,032 (0,075)
General (T-10)	-0,107 (0,076)
General (T-11)	0,010 (0,076)
General (T-12)	-0.049 (0.075)
General (T-13)	-0.137 (0.075)*
General (T-14)	-0.063 (0.075)
General (T-15)	-0.088 (0.074)
General (T-16)	0.024 (0.075)
General (T-17)	-0.073 (0.075)
General (T-18)	-0.004 (0.074)
General (T-19)	-0.062 (0.074)

General (T-20)	-0.076 (0.074)
General (T-21)	-0.055 (0.074)
General (T-22)	-0.074 (0.075)
General (T-23)	-0.056 (0.075)
General (T-24)	-0.131 (0.076)*
General (T-25)	-0.014 (0.076)
General (T-26)	-0.050 (0.076)
General (T-27)	-0.077 (0.076)
General (T-28)	0.013 (0.076)
General (T-29)	0.067 (0.076)
General (T-30)	-0.108 (0.077)
General (T-31)	-0.050 (0.077)
General (T-32)	0.160 (0.076)**
General (T-33)	0.092 (0.077)
General (T-34)	0.061 (0.078)
General (T-35)	0.071 (0.079)
General (T-36)	-0.013 (0.079)
General (T-37)	0.015 (0.081)
General (T-38)	-0.131 (0.081)
General (T-39)	-0.044 (0.083)
General (T-40)	-0.063 (0.079)
Symptoms (T-1)	0.038 (0.053)
Symptoms (T-2)	0.097 (0.054)*
Symptoms (T-3)	0.058 (0.054)
Symptoms (T-4)	0.042 (0.053)
Symptoms (T-5)	0.133 (0.053)**
Symptoms (T-6)	0.003 (0.052)
Symptoms (T-7)	0.031 (0.053)
Symptoms (T-8)	0.063 (0.053)
Symptoms (T-9)	0.079 (0.053)
Symptoms (T-10)	0.113 (0.053)**
Symptoms (T-11)	0.100 (0.052)*
Symptoms (T-12)	0.128 (0.052)**
Symptoms (T-13)	0.136 (0.052)***
Symptoms (T-14)	0.032 (0.051)
Symptoms (T-15)	0.043 (0.051)
Symptoms (T-16)	0.019 (0.051)
Symptoms (T-17)	0.018 (0.051)
Symptoms (T-18)	-0.027 (0.051)

Symptoms (T-19)	0.098 (0.051)*
Symptoms (T-20)	0.082 (0.050)
Symptoms (T-21)	0.133 (0.050)***
Symptoms (T-22)	0.086 (0.050)*
Symptoms (T-23)	0.143 (0.050)***
Symptoms (T-24)	0.130 (0.051)**
Symptoms (T-25)	0.111 (0.051)**
Symptoms (T-26)	0.121 (0.050)**
Symptoms (T-27)	0.111 (0.051)**
Symptoms (T-28)	0.076 (0.051)
Symptoms (T-29)	0.150 (0.051)***
Symptoms (T-30)	0.080 (0.051)
Symptoms (T-31)	0.110 (0.050)**
Symptoms (T-32)	0.110 (0.051)**
Symptoms (T-33)	0.076 (0.051)
Symptoms (T-34)	0.048 (0.051)
Symptoms (T-35)	-0.017 (0.051)
Symptoms (T-36)	-0.052 (0.051)
Symptoms (T-37)	-0.112 (0.051)**
Symptoms (T-38)	-0.077 (0.051)
Symptoms (T-39)	-0.074 (0.051)
Symptoms (T-40)	-0.016 (0.052)
Measures (T-1)	-0.040 (0.058)
Measures (T-2)	0.107 (0.059)*
Measures (T-3)	0.076 (0.058)
Measures (T-4)	0.036 (0.058)
Measures (T-5)	0.051 (0.057)
Measures (T-6)	0.095 (0.057)*
Measures (T-7)	0.043 (0.057)
Measures (T-8)	0.082 (0.058)
Measures (T-9)	0.023 (0.058)
Measures (T-10)	0.081 (0.058)
Measures (T-11)	0.023 (0.059)
Measures (T-12)	0.052 (0.058)
Measures (T-13)	0.170 (0.058)***
Measures (T-14)	0.027 (0.059)
Measures (T-15)	0.075 (0.058)
Measures (T-16)	0.030 (0.059)
Measures (T-17)	0.072 (0.059)

Measures (T-18)	-0.046 (0.060)
Measures (T-19)	-0.007 (0.060)
Measures (T-20)	-0.003 (0.059)
Measures (T-21)	0.004 (0.060)
Measures (T-22)	-0.036 (0.060)
Measures (T-23)	-0.066 (0.060)
Measures (T-24)	0.019 (0.060)
Measures (T-25)	-0.084 (0.060)
Measures (T-26)	-0.096 (0.061)
Measures (T-27)	-0.049 (0.061)
Measures (T-28)	-0.003 (0.062)
Measures (T-29)	-0.014 (0.062)
Measures (T-30)	-0.047 (0.063)
Measures (T-31)	-0.027 (0.064)
Measures (T-32)	-0.054 (0.064)
Measures (T-33)	0.013 (0.064)
Measures (T-34)	0.021 (0.065)
Measures (T-35)	0.060 (0.065)
Measures (T-36)	-0.018 (0.066)
Measures (T-37)	0.048 (0.067)
Measures (T-38)	0.109 (0.067)
Measures (T-39)	0.053 (0.066)
Measures (T-40)	0.046 (0.065)
Constant	-67.382 (18.994)***

Table 5: #Observations: 932. The standard errors are reported in parentheses. * p < 0.1; ** p < 0.05; *** p < 0.01.

Table 6. Regional fixed-effects model with the number of COVID-19 deaths as the dependent variable and Google searches as independent variables.

Deaths	**Coefficient (standard deviation)**
General	-0.021 (0.031)
Symptoms	-0.066 (0.022)***
Measures	0.022 (0.024)
General (T-1)	-0.021 (0.032)
General (T-2)	-0.021 (0.032)
General (T-3)	-0.033 (0.032)
General (T-4)	-0.017 (0.032)
General (T-5)	-0.038 (0.031)
General (T-6)	-0.017 (0.031)
General (T-7)	0.004 (0.031)

General (T-8)	0.003 (0.031)
General (T-9)	0.014 (0.031)
General (T-10)	-0.039 (0.031)
General (T-11)	0.006 (0.031)
General (T-12)	0.025 (0.031)
General (T-13)	-0.007 (0.031)
General (T-14)	-0.013 (0.031)
General (T-15)	-0.053 (0.031)*
General (T-16)	0.052 (0.031)*
General (T-17)	-0.033 (0.031)
General (T-18)	0.003 (0.031)
General (T-19)	-0.045 (0.031)
General (T-20)	0.010 (0.030)
General (T-21)	-0.038 (0.030)
General (T-22)	-0.017 (0.031)
General (T-23)	-0.008 (0.031)
General (T-24)	-0.023 (0.031)
General (T-25)	-0.005 (0.031)
General (T-26)	-0.010 (0.031)
General (T-27)	-0.051 (0.031)
General (T-28)	-0.039 (0.031)
General (T-29)	-0.008 (0.031)
General (T-30)	-0.015 (0.032)
General (T-31)	-0.041 (0.032)
General (T-32)	0.004 (0.031)
General (T-33)	-0.014 (0.032)
General (T-34)	0.003 (0.032)
General (T-35)	-0.018 (0.032)
General (T-36)	-0.005 (0.032)
General (T-37)	0.023 (0.033)
General (T-38)	-0.034 (0.034)
General (T-39)	-0.026 (0.034)
General (T-40)	-0.002 (0.032)
Symptoms (T-1)	-0.028 (0.022)
Symptoms (T-2)	0.005 (0.022)
Symptoms (T-3)	0.006 (0.022)
Symptoms (T-4)	0.013 (0.022)
Symptoms (T-5)	0.010 (0.022)
Symptoms (T-6)	-0.018 (0.022)

Symptoms (T-7)	-0.018 (0.022)
Symptoms (T-8)	0.007 (0.022)
Symptoms (T-9)	-0.005 (0.022)
Symptoms (T-10)	0.066 (0.022)***
Symptoms (T-11)	-0.004 (0.021)
Symptoms (T-12)	0.035 (0.021)
Symptoms (T-13)	0.023 (0.021)
Symptoms (T-14)	-0.005 (0.021)
Symptoms (T-15)	-0.008 (0.021)
Symptoms (T-16)	0.013 (0.021)
Symptoms (T-17)	0.028 (0.021)
Symptoms (T-18)	0.023 (0.021)
Symptoms (T-19)	0.040 (0.021)*
Symptoms (T-20)	0.023 (0.021)
Symptoms (T-21)	0.021 (0.021)
Symptoms (T-22)	0.001 (0.021)
Symptoms (T-23)	0.051 (0.021)**
Symptoms (T-24)	0.057 (0.021)***
Symptoms (T-25)	0.047 (0.021)**
Symptoms (T-26)	0.046 (0.021)**
Symptoms (T-27)	0.069 (0.021)***
Symptoms (T-28)	0.047 (0.021)**
Symptoms (T-29)	0.036 (0.021)*
Symptoms (T-30)	0.078 (0.021)***
Symptoms (T-31)	0.052 (0.021)**
Symptoms (T-32)	0.048 (0.021)**
Symptoms (T-33)	0.043 (0.021)**
Symptoms (T-34)	0.025 (0.021)
Symptoms (T-35)	0.028 (0.021)
Symptoms (T-36)	0.014 (0.021)
Symptoms (T-37)	0.024 (0.021)
Symptoms (T-38)	0.053 (0.021)**
Symptoms (T-39)	-0.009 (0.021)
Symptoms (T-40)	0.014 (0.021)
Measures (T-1)	0.036 (0.024)
Measures (T-2)	-0.010 (0.024)
Measures (T-3)	0.005 (0.024)
Measures (T-4)	-0.003 (0.024)
Measures (T-5)	-0.002 (0.024)

Measures (T-6)	0.027 (0.024)
Measures (T-7)	0.027 (0.024)
Measures (T-8)	0.009 (0.024)
Measures (T-9)	0.027 (0.024)
Measures (T-10)	0.002 (0.024)
Measures (T-11)	0.001 (0.024)
Measures (T-12)	0.006 (0.024)
Measures (T-13)	0.012 (0.024)
Measures (T-14)	0.016 (0.024)
Measures (T-15)	0.061 (0.024)**
Measures (T-16)	0.040 (0.024)*
Measures (T-17)	0.004 (0.024)
Measures (T-18)	-0.003 (0.024)
Measures (T-19)	-0.011 (0.025)
Measures (T-20)	-0.011 (0.025)
Measures (T-21)	0.010 (0.025)
Measures (T-22)	0.033 (0.025)
Measures (T-23)	0.014 (0.025)
Measures (T-24)	-0.011 (0.025)
Measures (T-25)	-0.020 (0.025)
Measures (T-26)	0.012 (0.025)
Measures (T-27)	-0.047 (0.025)*
Measures (T-28)	-0.024 (0.025)
Measures (T-29)	0.025 (0.026)
Measures (T-30)	-0.023 (0.026)
Measures (T-31)	0.008 (0.026)
Measures (T-32)	-0.010 (0.026)
Measures (T-33)	0.011 (0.026)
Measures (T-34)	-0.004 (0.027)
Measures (T-35)	-0.026 (0.027)
Measures (T-36)	0.002 (0.027)
Measures (T-37)	0.013 (0.027)
Measures (T-38)	-0.014 (0.027)
Measures (T-39)	-0.012 (0.027)
Measures (T-40)	0.033 (0.027)
Constant	-20.185 (7.856)***

Table 6: #Observations: 946. The standard errors are reported in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Appendix B

Table 7. Regional fixed-effects model with the number of COVID-19 cases as the dependent variable and Google searches as independent variables with maximally 32 lags.

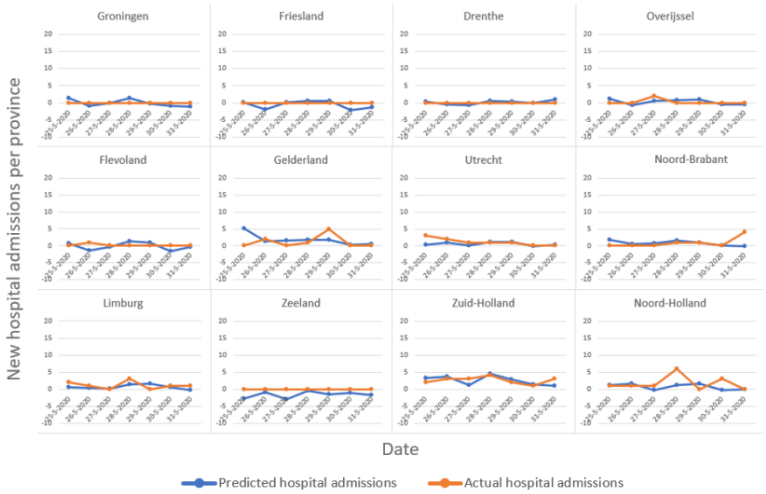
Cases	Coefficient (standard deviation)
General	-0.016 (0.119)
General (T-1)	-0.110 (0.123)
General (T-2)	0.059 (0.120)
General (T-4)	0.023 (0.112)
General (T-6)	0.009 (0.108)
General (T-9)	-0.177 (0.111)
General (T-11)	-0.106 (0.118)
General (T-12)	-0.125 (0.122)
General (T-13)	0.012 (0.117)
General (T-15)	-0.177 (0.118)
General (T-16)	-0.141 (0.126)
General (T-17)	-0.324 (0.120)***
General (T-19)	-0.197 (0.121)
General (T-20)	-0.220 (0.126)*
General (T-21)	-0.165 (0.125)
General (T-22)	-0.380 (0.118)****
General (T-30)	-0.090 (0.122)
General (T-31)	-0.052 (0.123)
Symptoms	0.007 (0.085)
Symptoms (T-5)	0.132 (0.084)
Symptoms (T-8)	0.066 (0.084)
Symptoms (T-10)	0.172 (0.085)**
Symptoms (T-11)	0.202 (0.085)**
Symptoms (T-12)	0.201 (0.085)**
Symptoms (T-13)	0.140 (0.085)*
Symptoms (T-14)	0.190 (0.084)**
Symptoms (T-16)	0.194 (0.085)**
Symptoms (T-17)	0.286 (0.085)***
Symptoms (T-18)	0.264 (0.086)***
Symptoms (T-19)	0.269 (0.085)***
Symptoms (T-20)	0.228 (0.085)***

Symptoms (T-21)	0.244 (0.085)***
Symptoms (T-22)	0.269 (0.084)***
Symptoms (T-23)	0.163 (0.084)*
Symptoms (T-24)	0.286 (0.085)***
Symptoms (T-25)	0.267 (0.084)***
Symptoms (T-26)	0.160 (0.084)*
Symptoms (T-27)	0.245 (0.084)***
Symptoms (T-28)	0.174 (0.085)**
Symptoms (T-29)	0.210 (0.085)**
Symptoms (T-30)	0.178 (0.084)**
Symptoms (T-31)	0.215 (0.084)**
Symptoms (T-32)	0.214 (0.084)**
Measures	0.104 (0.092)
Measures (T-1)	0.279 (0.094)***
Measures (T-2)	0.244 (0.095)***
Measures (T-3)	0.101 (0.097)
Measures (T-4)	0.120 (0.097)
Measures (T-5)	0.051 (0.096)
Measures (T-7)	0.106 (0.096)
Measures (T-8)	0.122 (0.098)
Measures (T-9)	0.161 (0.100)
Measures (T-10)	0.323 (0.099)***
Measures (T-11)	0.142 (0.099)
Measures (T-14)	0.103 (0.099)
Measures (T-16)	0.117 (0.101)
Measures (T-17)	0.166 (0.100)*
Measures (T-18)	0.050 (0.100)
Measures (T-20)	0.208 (0.099)**
Measures (T-24)	0.101 (0.095)
Measures (T-26)	-0.111 (0.098)
Measures (T-30)	0.098 (0.106)
Measures (T-31)	0.310 (0.107)***
Measures (T-32)	0.169 (0.101)*
Constant	-258.182 (14.361)***

Table 7: #Observations: 1,130. Model specification is based on adaptive lasso-estimates. The standard errors are reported in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

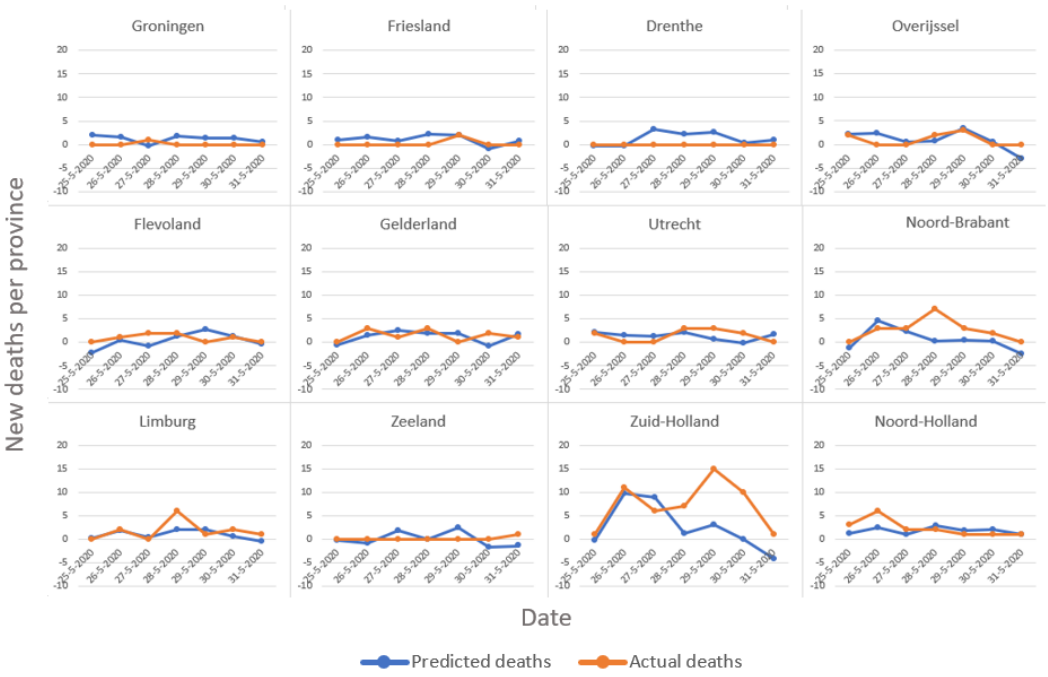
Appendix C

Graph 3: Actual number of hospital admissions per day per province compared to predicted number of hospital admissions per day per province.



Graph 3: #Observations: 153. Prediction based on plugin lasso-estimates. R^2: 0.600; MSE: 36.35. Average number of hospital admissions per day in last week of May: 0.76.

Graph 4: Actual number of COVID-19 deaths per day per province compared to predicted number of COVID-19 deaths per day per province



Graph 4: #Observations: 128. Prediction based on plugin lasso-estimates. R^2: 0.627; MSE: 23.86. Average number of deaths per day in last week of May: 1.75.

Bibliography

Benigeri, M., & Pluye, P. (2003). Shortcomings of health information on the Internet. Health promotion international, 18(4), 381-386.

Brunori, P., & Resce, G. (2020). Searching for the peak Google Trends and the Covid-19 outbreak in Italy.

Carneiro, H. A., & Mylonakis, E. (2009). Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clinical infectious diseases*, *49*(10), 1557-1564.

Centraal Bureau voor de Statistiek. (2020, April 24). Sterfte neemt af in derde week van april 2020. Retrieved from https://www.cbs.nl/nl-nl/nieuws/2020/17/sterfte-neemt-af-in-derde-week-van-april-2020.

Farzanegan, M. R., Feizi, M., & Sadati, S. M. (2020). *Google It Up! A Google Trends-based analysis of COVID-19 outbreak in Iran* (No. 17-2020). Joint Discussion Paper Series in Economics.

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*(7232), 1012-1014.

Hu, D., Lou, X., Xu, Z., Meng, N., Xie, Q., Zhang, M., ... & Wang, F. (2020). More effective strategies are required to strengthen public awareness of COVID-19: Evidence from Google Trends. *Journal of Global Health*, *10*(1).

Husnayain, A., Fuad, A., & Su, E. C. Y. (2020). Applications of google search trends for risk communication in infectious disease management: A case study of COVID-19 outbreak in Taiwan. *International Journal of Infectious Diseases*.

Li, C., Chen, L. J., Chen, X., Zhang, M., Pang, C. P., & Chen, H. (2020). Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020. *Eurosurveillance*, *25*(10), 2000199.

Liu, Y., Gayle, A. A., Wilder-Smith, A., & Rocklöv, J. (2020). The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of travel medicine.*

McMullan, M. (2006). Patients using the Internet to obtain health information: how this affects the patient–health professional relationship. Patient education and counseling, 63(1-2), 24-28.

Morsy, S., Dang, T. N., Kamel, M. G., Zayan, A. H., Makram, O. M., Elhady, M., ... & Huy, N. T. (2018). Prediction of Zika-confirmed cases in Brazil and Colombia using Google Trends. *Epidemiology & Infection*, *146*(13), 1625-1627.

Rijksinstituut voor Volksgezondheid en Milieu. (2020). Actuele informatie over het nieuwe coronavirus (COVID-19). Retrieved from https://www.rivm.nl/coronavirus-covid-19/actueel.

Sciascia, S., & Radin, M. (2017). What can Google and Wikipedia can tell us about a disease? Big Data trends analysis in Systemic Lupus Erythematosus. *International journal of medical informatics*, *107*, 65-69.

Surveillances, V. (2020). The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19)—China, 2020. *China CDC Weekly*, *2*(8), 113-122.

Suziedelyte, A. (2012). How does searching for health information on the Internet affect individuals' demand for health care services?. Social science & medicine, 75(10), 1828-1835.

Tan, S. S. L., & Goonawardene, N. (2017). Internet health information seeking and the patient-physician relationship: a systematic review. Journal of medical Internet research, 19(1), e9.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

Tonsaker, T., Bartlett, G., & Trpkov, C. (2014). Health information on the Internet: gold mine or minefield?. Canadian Family Physician, 60(5), 407-408.

World Health Organization. (2020c, April 1). Coronavirus disease 2019 (COVID-19) Situation Report – 72. Retrieved from https://apps.who.int/iris/bitstream/handle/10665/331685/nCoVsitrep01Apr2020-eng.pdf

World Health Organization. (2020c, April 2). Coronavirus disease 2019 (COVID-19) Situation Report – 73. Retrieved from https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200402-sitrep-73-covid-19.pdf?sfvrsn=5ae25bc7_4

World Health Organization. (2020). Naming the coronavirus disease (COVID-19) and the virus that causes it. Retrieved from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it

World Health Organization. (2020b, February 16). Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19) . Retrieved from https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf