

Google Trends reflect allergic rhinitis symptoms related to birch and grass pollen seasons

Kostas Karatzas  · Lazaros Papamanolis · Nikos Katsifarakis · Marina Riga · Barbora Werchan · Matthias Werchan · Uwe Berger · Karl-Christian Bergmann

Received: 2 June 2018 / Accepted: 6 September 2018
© Springer Nature B.V. 2018

Abstract Google Trends (GT) describes the variation of the relevant interest of internet searches toward medical conditions and related symptoms. Allergic rhinitis symptom levels result from the intensity of exposure to aeroallergens in combination with relevant medication use. We analyze data from Germany to examine the relationship between hay fever-related Google search terms, symptom levels, medication use, and pollen count levels. For doing so, we also employ the new definitions on pollen season and peak pollen

period start and end as proposed by the European Academy of Allergy and Clinical Immunology in a recently published position paper. We extract GT data for a number of search terms related to allergic rhinitis for Germany. We use total nasal symptom and mediation scores as reported by patients via a patient hay fever diary in the Berlin and Brandenburg areas in Germany for 3 years (2014–2016), accompanied by pollen data. Then a Pearson and Spearman correlation analysis is performed between symptom data and GT

K. Karatzas (✉) · L. Papamanolis · N. Katsifarakis · M. Riga
Environmental Informatics Research Group, Department of Mechanical Engineering, Aristotle University, University Campus, Box 483, 54124 Thessaloniki, Greece
e-mail: kkara@eng.auth.gr

L. Papamanolis
e-mail: lpapamanolis@yahoo.com

N. Katsifarakis
e-mail: nikolakk@isag.meng.auth.gr

M. Riga
e-mail: mriga@isag.meng.auth.gr

B. Werchan · M. Werchan · K.-C. Bergmann
German Pollen Information Service Foundation, Berlin, Germany
e-mail: barbora.werchan@charite.de

M. Werchan
e-mail: matthias.werchan@charite.de

K.-C. Bergmann
e-mail: karlchristianbergmann@gmail.com

B. Werchan · M. Werchan · K.-C. Bergmann
Charité – Universitätsmedizin Berlin, Department of Dermatology, Venerology and Allergology, Berlin, Germany

B. Werchan · M. Werchan · K.-C. Bergmann
Freie Universität Berlin, Berlin, Germany

B. Werchan · M. Werchan · K.-C. Bergmann
Humboldt-Universität zu Berlin, Berlin, Germany

B. Werchan · M. Werchan · K.-C. Bergmann
Department of Dermatology, Venerology and Allergology, Berlin Institute of Health, Berlin, Germany

U. Berger
Aerobiology and Pollen Information Research Unit, Department of Oto-Rhino-Laryngology, Medical University of Vienna, Vienna, Austria
e-mail: uwe.berger@meduniwien.ac.at

data. A graphical analysis is conducted, and the identification of pollen season and peak pollen periods is done based on the EAACI criteria. The analysis reveals that GT data are highly correlated with symptom levels and follow peak pollen period start–end, concerning grass and birch pollen-induced allergic rhinitis symptoms. GT data can be used as a proxy for the identification of the onset and variation of nasal symptom and medication score for allergic rhinitis sufferers.

Keywords Google Trends · Allergic rhinitis · Pollen counts · Pollen season · Symptom variation

1 Introduction

The use of data volumes resulting from internet search, as an indicator of actual environmental or medical events and problems, has increased in the last years (Ling and Lee 2016; Riga et al. 2015; Tkachenko et al. 2017). One of these types of data result from Google Trends (GT, www.google.com/Trends) which is “a public web facility of Google Inc., based on Google Search, that shows how often a particular search-term is entered relative to the total search volume across various regions of the world, and in various languages” (Wikipedia 2018). These data have already been used for the investigation of (a) the interest of the general public as well as of allergy sufferers on aeroallergens and related symptoms and (b) the relationship between GT data and pollen count levels (Bousquet et al. 2017; Kang et al. 2015; Mösges et al. 2011; Oteros et al. 2015). In the frame of this study, we move further to investigate the relationship between GT-based interest and pollen-related symptoms reported by sufferers (Bastl et al. 2014; Karatzas et al. 2014). This work follows the recently published European Academy of Allergy and Clinical Immunology (EAACI) position paper on the definition of pollen seasons (Pfaar et al. 2017) as well as the recently proved mapping between aforementioned criteria and actual symptom levels reported by patients on a voluntary basis (Karatzas et al. 2018).

2 Materials and methods

2.1 Pollen, symptom, medication, and Google Trend data used

The pollen count level data as well as the pollen-induced symptom and medication use data came from the Berlin and Brandenburg areas situated in the northwest of Germany. Four pollen traps provided the pollen monitoring information in Berlin, Neustrelitz, Potsdam, and Treuenbrietzen (Fig. 1). The species monitored included birch (*Betula*) and grass (Poaceae), with concentration values per day recorded for the years 2014, 2015, and 2016. Values range from 9 to 3042 pollen grains/m³ for birch and between 1 and 173 pollen grains/m³ for grass. Symptom and medication use data resulted from patient reports were collected on a voluntary basis with the aid of the patient’s hay fever diary (PHD) on a daily basis (Berger et al. 2011). The term patients here corresponds to those reporting symptoms via the PHD, which are not patients by definition (i.e. with an allergy being diagnosed), as anybody can report symptoms through the system. In this work, we make use of the total nasal symptom and medication score (TNSMS) which ranges from 0 to 9.05 as explained in detail in Bastl et al. (2014) and in Karatzas et al. (2014).

We utilize GT services to extract data concerning the volume of Google searches in Germany, making use of specific aeroallergen-related keywords, as those were specified by allergology experts and are given in Table 1, in a way similar to the one already published in the related literature (Bousquet et al. 2017; Mösges et al. 2011). In order to cope with the pollen season, we focused our GT analysis on the period starting on the 1 April (for 2014 and 2015) and 31 March (for 2016 due to leap year) and for 150 consecutive days (i.e., till the end of August), for each one of the 3 years of study (Karatzas et al. 2018). GT data present the relevant frequency of the number of searches conducted via the Google search engine for a specific country and for the related time interval, thus ranging from 0 to 100. The GT data that we found for Germany were of better quality in terms of completeness and time period representation in comparison with data for the Berlin and Brandenburg areas alone. For this reason, we considered them to represent average Google search conditions in Germany overall, and thus, they were rendered as representative aggregates concerning any

Fig. 1 Location of the four pollen monitoring stations used in the Berlin–Brandenburg areas, Germany (map source: www.openstreetmap.org, 2018)



Table 1 Keywords used for the Google Trends analysis in Germany

German keyword	Equivalent English keyword
Allergie	Allergy
Asthma	Asthma
Bindehautentzündung	Conjunctivitis
Birkenpollen	Birch pollen
Gräserpollen	Grass pollen
Heuschnupfen	Hay fever
Pollen	Pollen
Pollenflug	Pollen count
Pollenvorhersage	Pollen forecast

region of the country like Berlin and Brandenburg areas from where our pollen count data came from. This approach is also justified by the fact that the symptom data used from this specific geographic area are actually the mean TNSMS value per day from all those PHD users reporting their symptoms. This means that an average aggregated symptom score of the geographic region of interest is compared with GT data calculated (and aggregated) at a country level.

The identification of the pollen season (PS) and of the peak pollen periods (PPP) per calendar year was based on the relevant EAACI criteria (Pfaar et al. 2017).

2.2 Correlation analysis

We use the symptom and medication use data in order to calculate the Pearson correlation coefficient between GT and the TNSMS according to the following formulae:

$$r = \frac{\sum_{i=1}^N (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}} \quad (1)$$

where x represents the TNSMS, y represents GT levels, and N represents the number of available data records per year of study.

The Pearson correlation coefficient expresses the strength of the association (in terms of the degree of linear dependency) between the studied parameters. In order to decide whether the resulting correlation occurred by chance, we run a test of significance and calculate the p value for each correlation coefficient. This test assumes that studied variables follow the normal distribution, while both variables are randomly sampled (Campbell et al. 2007). If the resulting p values are higher than the level of significance

chosen (here 5% or 0.05), then we consider the relevant correlation coefficient to be statistically insignificant within this confidence interval.

In order to address the option where studied variables are not normally distributed, we employ the Spearman correlation coefficient, which describes how well a monotonic function can depict the relationship between the studied parameters:

$$r_s = 1 - \frac{6 \sum_{i=1}^N D_i}{n(n^2 - 1)}. \quad (2)$$

Here D_i represents the difference between the two sets of data, $\text{ranked}(x)$ and $\text{ranked}(y)$, ranked denoting that both datasets are ranked from higher to lower values (Gibbons and Chakraborti 2011), and x and y represent TNSMS and GT data, respectively. We proceed and we calculate again the p values for a significance level of $\alpha = 0.05$ to check for the statistical significance of the correlation coefficient.

As we wanted to check whether pollen levels and pollen season definitions are reflected on patient symptoms, we created relevant plots covering the time period of study for each one of the 3 years under

Table 2 Pearson (up) and Spearman (down) correlation coefficient and p value between the TNSMS and GT search terms, calculated for the years 2014, 2015, and 2016 for a 150-day period starting on 1 April

Year	2014		2015		2016	
	r	p	r	p	r	p
Pearson and p levels						
Allergy	0.78	3.09E−32	0.79	9.15E−34	0.82	2.31E−38
Asthma	0.48	2.77E−10	0.44	4.4E−08	0.44	1.5E−08
Conjunctivitis	0.48	4.41E−10	0.39	5.54E−07	0.61	7.16E−17
Birch pollen	0.61	1.43E−16	0.38	1.01E−06	0.60	5.92E−16
Grass pollen	0.27	0.000706	0.50	6.72E−11	0.44	1.81E−08
Hay fever	0.83	1.49E−39	0.86	4.87E−45	0.84	1.7E−42
Pollen	0.75	1.02E−28	0.76	2.04E−30	0.80	1.99E−34
Pollen count	0.78	8.42E−33	0.78	1.41E−32	0.78	5.29E−32
Pollen forecast	0.35	1.06E−05	0.52	6.06E−12	0.41	1.2E−07
Spearman and p levels						
Allergy	0.66	4.29E−20	0.79	6.98E−34	0.85	3.83E−43
Asthma	0.26	1.08E−03	0.41	1.85E−07	0.45	6.60E−09
Conjunctivitis	0.46	2.32E−09	0.36	4.53E−06	0.63	1.88E−18
Birch pollen	0.32	5.15E−05	0.46	3.78E−09	0.53	1.42E−12
Grass pollen	0.22	0.007072	0.47	1.09E−09	0.45	8.86E−09
Hay fever	0.72	2.10E−25	0.88	1.24E−49	0.93	1.01E−65
Pollen	0.64	5.56E−19	0.80	2.19E−34	0.86	9.48E−46
Pollen count	0.68	4.03E−22	0.83	5.88E−40	0.85	9.05E−43
Pollen forecast	0.28	3.91E−04	0.55	1.86E−13	0.41	1.22E−07

Highest correlation values are indicated with bold

investigation. The differences in the descriptive statistics of the parameters of interest (as given in Table 2) led to a data normalization step before plotting; for each one of the initial parameters of interest, a normalized parameter was estimated and calculated as follows:

$$z' = \frac{z - \bar{z}}{\sigma} \quad (3)$$

where $\bar{z}$ is the mean value and σ is the standard deviation for each one of the pollen and symptom data included in the analysis.

Moreover, we conducted detailed plots per year for some of the GT data in parallel with pollen and symptom data, with the aim to investigate their temporal variation, relationship, and correlation. All calculations and analyses were performed with the aid of Matlab R2017a (www.mathworks.com).

3 Results

The Pearson correlation coefficient between the keywords used in the GT analysis for Germany and the

TNSMS as recorded via the PHD users for Berlin and Brandenburg areas suggests that correlation levels vary among keywords; there are keywords that score very high for all 3 years of study, the highest of all achieved by *Heuschnupfen* (hay fever) with r ranging between 0.83 and 0.86. The terms *Allergie* (allergy), *Pollenflug* (pollen count), and *Pollen* (pollen) present also with a high correlation coefficient, higher than 0.75 in all cases. The correlation coefficient levels are found to be significant in terms of p value (0.05).

For Spearman correlation coefficient, *Heuschnupfen* (hay fever) is again the GT keyword that demonstrates the higher value ranging from 0.72 to 0.93. Taking into account the p value, all correlation coefficients may again be considered being significant.

The results obtained lead to the decision to select the term *Heuschnupfen* (hay fever) for a further graphical investigation (Figs. 2, 3, 4). The aforementioned figures also present pollen season as well as peak pollen days in parallel with pollen count and TNSMS data. It is evident that in all 3 years of study, GT data reflect symptom data. Moreover, peaks in GT data appear in three out of the four PPP dates in 2014

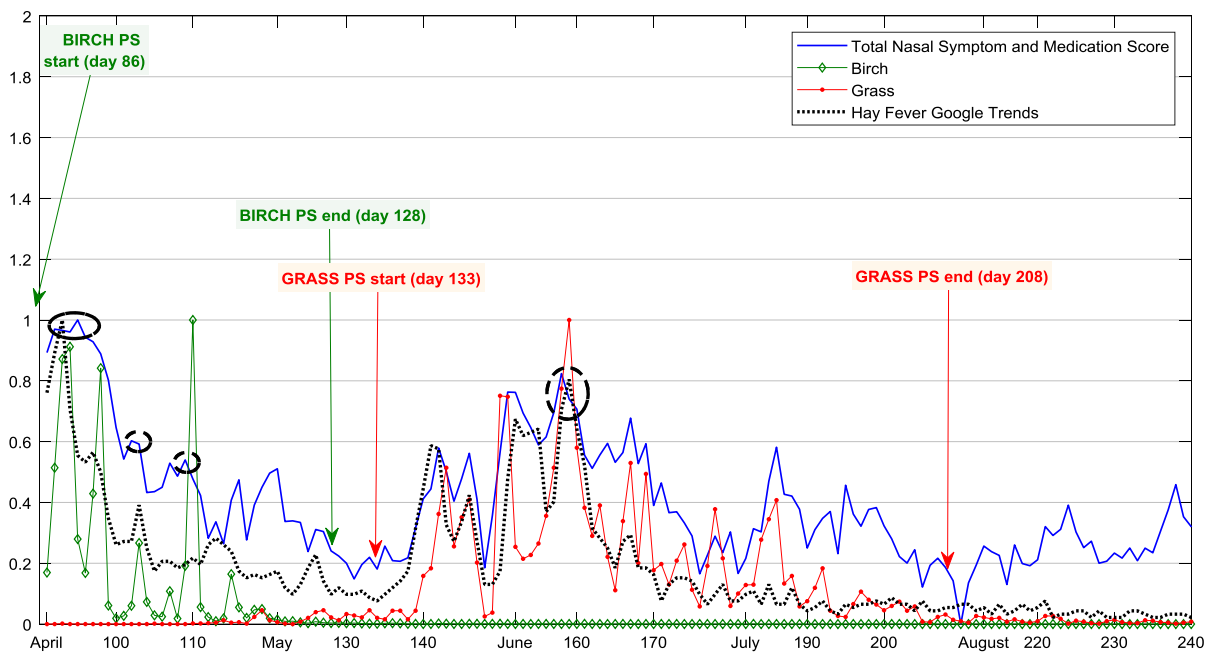


Fig. 2 Temporal variation of the total nasal symptom and medication score (TNSMS) as well as birch and grass pollen levels from 1 April and for 150 consecutive days in 2014 for the Berlin and Brandenburg areas. Google Trends data are aggregated and correspond to the whole of Germany. The

horizontal axis reports month start and calendar day number. The pollen season start and end days are marked with relevant arrows with the related numbered calendar being mentioned. Heavy dashed lines represent high pollen periods according to the EAACI criteria

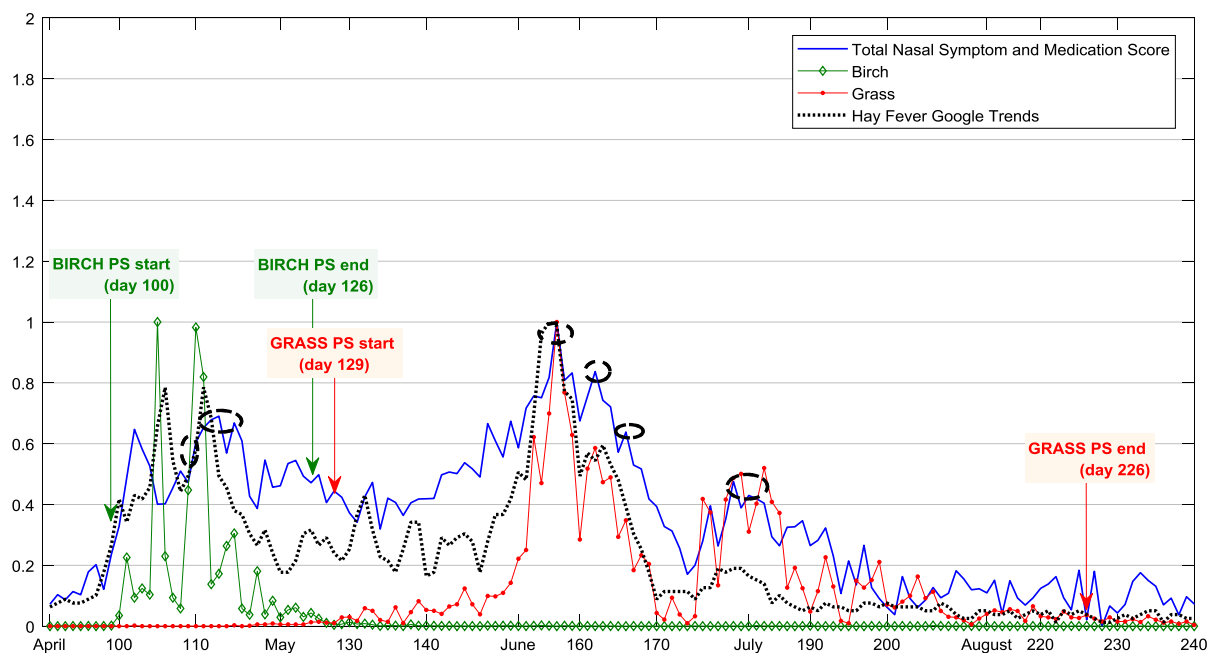


Fig. 3 Temporal variation of the total nasal symptom and medication score (TNSMS) as well as birch and grass pollen levels from 1 April and for 150 consecutive days in 2015 for the Berlin and Brandenburg areas. Google Trends data are aggregated and correspond to the whole of Germany. The

horizontal axis reports month start and calendar day number. The pollen season start and end days are marked with relevant arrows with the related numbered calendar being mentioned. Heavy dashed lines represent high pollen periods according to the EAACI criteria

and in four out of five PPP dates in 2015 and 2016. In addition and in all 3 years, GT data present with their first pick just before the TNSMS also demonstrates its first peak of the season (with a lag of 1–2 days).

4 Discussion

The Pearson and the Spearman correlation coefficients between GT data for Germany and the TNSMS for the Berlin and Brandenburg areas are very high. “Hay fever” (*Heuschnupfen*) proved to be the GT term with the highest correlation with symptoms. The graphical investigation of GT data for this keyword reveals that it very well follows the TNSMS onset and variation for all 3 years of study. (It should be noted that the TNSMS data collected with the aid of the PHD are used here as a “proxy” to understand symptoms in the whole population.) Moreover, GT data demonstrate peaks exactly when expected on the basis of the EAACI criteria for PPP. In addition, GT data precede TNSMS data in most cases at the beginning of both birch and grass pollen seasons, thus indicating that

interested people are either aware or they experience an early reaction to the increasing pollen levels that will follow. Aforementioned findings are in line with previous studies for Germany (König and Mösges 2014) and support the use of GT data for identifying the temporal variation as well as the onset of rhinitis symptoms.

5 Conclusions

It has been suggested that GT data are an expression of the actual epidemiology of allergic rhinitis symptoms and may be used as an additional means to study the development of the disease and its symptoms (Kang et al. 2015). In this study, we show that GT data can be used as a proxy for the development of allergic rhinitis symptoms and more specifically for identifying the onset, the variation, and the levels of the TNSMS for hay fever sufferers. Our study also demonstrates that GT data reflect the pollen season as resulting from the EAACI definition (Pfaar et al. 2017).

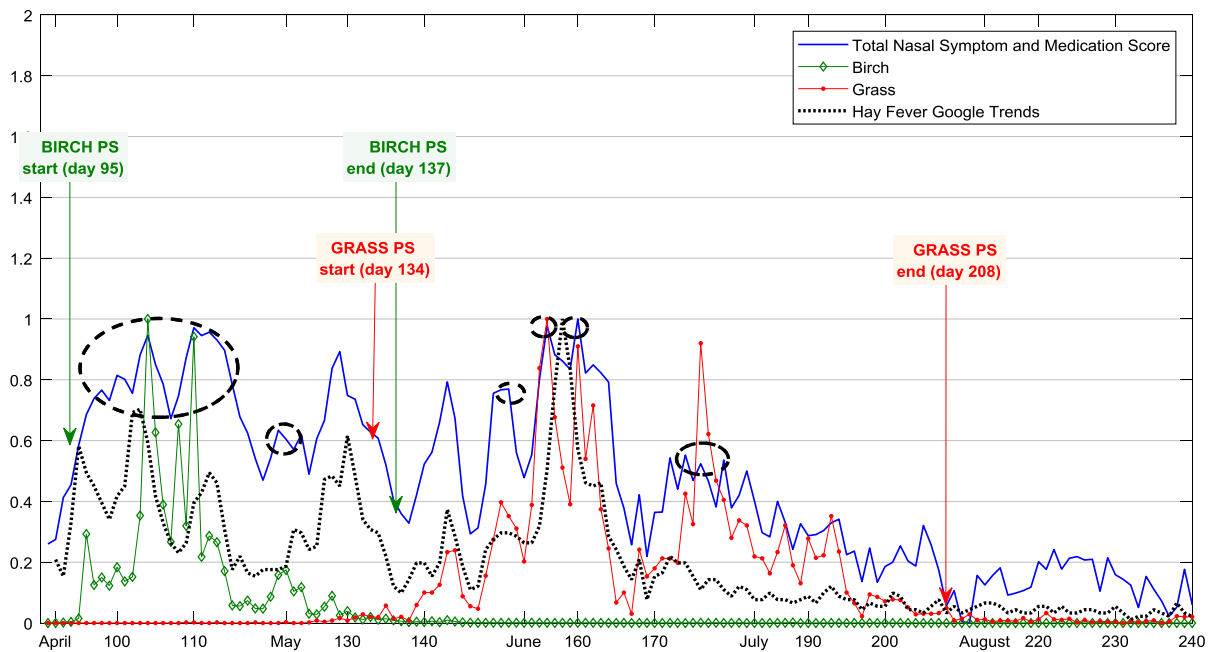


Fig. 4 Temporal variation of the total nasal symptom and medication score (TNSMS) as well as birch and grass pollen levels from 31 March and for 150 consecutive days in 2016 for the Berlin and Brandenburg areas. Google Trends data are aggregated and correspond to the whole of Germany. The

horizontal axis reports month start and calendar day number. The pollen season start and end days are marked with relevant arrows with the related numbered calendar being mentioned. Heavy dashed lines represent high pollen periods according to the EAACI criteria

On this basis, it is evident that clinical trials of allergen immunotherapy for pollen-induced rhinoconjunctivitis may also benefit from GT data analytics. Future research will focus on the development of mathematical models that would be able to describe pollen-induced symptoms and GT data for supporting pollen forecast information for the general public.

Author's contribution KCB and KK conceived and initiated the study, LP prepared the Google Trends data, BW and MW prepared the pollen and the symptom data, KK designed the study framework, KK, NK, and MR made the computations and graphical analysis, KK drafted the manuscript, UB provided with expert support, and KCB provided with expert advice and support.

Compliance with ethical standards

Conflict of interest The authors declare that they have no competing interests

References

- Bastl, K., Kmenta, M., Jäger, S., Bergmann, K. C., & Berger, U. (2014). Development of a symptom load index: Enabling temporal and regional pollen season comparisons and pointing out the need for personalized pollen information. *Aerobiologia*, 30(3), 269–280.
- Berger, U., Jäger, S., & Bergmann, K. C. (2011). PHD, the electronic patient's hayfever diary. *European Respiratory Journal*, 38(Suppl 55), 3213.
- Bousquet, J., Agache, I., Anto, J. M., Bergmann, K. C., Bachert, C., Annesi-Maesano, I., et al. (2017). Google Trends terms reporting rhinitis and related topics differ in European countries. *Allergy*, 72(8), 1261–1266.
- Campbell, M. J., Machin, D., & Walters, S. J. (2007). *Medical statistics: A textbook for the health sciences* (4th ed.). Chichester: Wiley.
- Gibbons, J. D., & Chakraborti, S. (2011). *Nonparametric statistical inference*. Florida: Chapman & Hall.
- Kang, M. G., Song, W. J., Choi, S., Kim, H., Ha, H., Kim, S. H., et al. (2015). Google unveils a glimpse of allergic rhinitis in the real world. *Allergy*, 70(1), 124–128.
- Karatzas, K., Katsifarakis, N., Riga, M., Werchan, B., Werchan, M., Berger, U., et al. (2018). New European Academy of Allergy and Clinical Immunology definition on pollen season mirrors symptom load for grass and birch pollen-induced allergic rhinitis. *Allergy*, 73, 851–1859. <https://doi.org/10.1111/all.13487>.

- Karatzas, K., Voukantsis, D., Jäger, S., Berger, U., Smith, M., Brandt, O., et al. (2014). The patient's hay-fever diary: Three-years of results from Germany. *Aerobiologia*, 30(1), 1–11.
- König, V., & Mösges, R. (2014). A model for the determination of pollen count using Google search queries for patients suffering from allergic rhinitis. *Journal of Allergy (Cairo)*, 2014, 381983.
- Ling, R., & Lee, J. (2016). Disease monitoring and health campaign evaluation using Google Search activities for HIV and AIDS, stroke, colorectal cancer, and marijuana use in Canada: A retrospective observational study. *JMIR Public Health and Surveillance*, 2(2), e156.
- Mösges, R., Adrian, M., el Hassan, E., & König, V. (2011). What Google® knows about the pollen season. *Allergy*, 66(5), 707–708.
- Oteros, J., García-Mozo, H., Viuf Ørby, P., & Galán, C. (2015). Google Trends, useful tool in airborne pollen detection. *Anales de la Facultad de Medicina*, 76(3), 265–268.
- Pfaar, O., Bastl, K., Berger, U., Buters, J., Calderon, M. A., Clot, B., et al. (2017). Defining pollen exposure times for clinical trials of allergen immunotherapy for pollen-induced rhinoconjunctivitis—An EAACI position paper. *Allergy*, 72(5), 713–722.
- Riga, M., Stocker, M., Rönkkö, M., Karatzas, K., & Kolehmainen, M. (2015). Air quality information extraction from Twitter with the use of Self-Organizing Maps. *Journal of Environmental Informatics*, 26(1), 27–40.
- Tkachenko, N., Chotvijit, S., Gupta, N., Bradley, E., Gilks, C., Guo, W., et al. (2017). Google Trends can improve surveillance of Type 2 diabetes. *Scientific Reports*, 7, 4993.
- Wikipedia. (2018). *Google Trends*. https://en.wikipedia.org/wiki/Google_Trends. Accessed 1 June 2018.