



OPEN

## Cancer incidence and digital information seeking in Germany: a retrospective observational study

Hannah Wecker<sup>1</sup>, Daniel Maier<sup>2,3</sup>, Stefanie Ziehfreund<sup>1</sup>, Fabienne A. U. Fox<sup>2</sup>, Ian Erhard<sup>2,3</sup>, Jörg Janne Vehreschild<sup>2,4,5</sup> & Alexander Zink<sup>1,6</sup>✉

Awareness is vital for cancer prevention. US studies show a strong link between web searches and cancer incidence. In Europe, the relationship remains unclear. This study characterizes regional and temporal relationships between cancer incidence and web searches and investigates the content of searches related to breast, cervical, colorectal, lung, prostate, and testicular cancer, brain tumors, and melanoma in Germany (July 2018–December 2019). Aggregate data from Google Ads Keyword Planner and national cancer registry data were analyzed. Spearman's correlation coefficient ( $r_s$ ) examined associations between cancer incidence and web search, repeated measures correlation ( $r_{rm}$ ) assessed time trends and searches were qualitatively categorized. The frequency of malignancy-related web searches correlated with cancer incidence ( $r_s = 0.88$ ,  $P = 0.007$ ), e.g., breast cancer had more queries than the lower-incidence cervical cancer. Seasonally, incidence and searches followed similar patterns, peaking in spring and fall, except for melanoma. Correlations between entity incidence and searches ( $0.037 \leq r_{rm} \leq 0.208$ ) varied regionally. Keywords mainly focused on *diagnosis*, *symptoms*, and *general information*, with variations between entities. In Germany, web searches correlated with regional and seasonal incidence, revealing differences between North/East and South/West. These insights may help improve prevention strategies by identifying regional needs and assessing impact of awareness campaigns.

**Keywords** Digital epidemiology, Infodemiology, Cancer prevention, Pan-cancer, Web search behavior

Cancer is the second most frequent cause of death in Europe<sup>1</sup>. In Germany, more than 498,000 patients were diagnosed with cancer in 2018, with the most prevalent entities including breast, prostate, colorectal, lung, and skin cancer<sup>2</sup>. About 37% of these cancer diagnoses are estimated to have been preventable<sup>3–6</sup>.

Internet use is steadily rising alongside the rapid growth of information, including health-related materials. Individuals frequently turn to search engines to inform themselves about diseases and assess their own condition<sup>7</sup>. In particular, patients with a diagnosed disease and their care providers search for information about respective diagnoses, risk factors, associated symptoms, prognosis, and therapy options<sup>8–13</sup>. This behavior applies to cancer as well<sup>14</sup>. One survey reported that approximately two-thirds of cancer patients turn to the internet in search of cancer-related information<sup>15</sup>.

Web search engines record user search history and location. These data can be analyzed to evaluate web search patterns, create search profiles, and infer user interests<sup>16–18</sup>. Studies from the US have found associations between web search volume (SV), the number of search queries for a specific search term, and the incidence of various cancer types<sup>19,20</sup>. Moreover, disease-specific web searches have been observed to spike after disease prevention campaigns, especially with respect to the Breast Cancer Awareness Month<sup>18,21–23</sup> and after media coverage of diseased celebrities<sup>24–26</sup>.

While this body of research suggests that SV could be used as a proxy for public disease awareness<sup>18,19,21</sup>, it must be highlighted that SV as a type of digital trace data<sup>27</sup> cannot be equated with the complex construct of

<sup>1</sup>Department of Dermatology and Allergy, TUM School of Medicine and Health, Technical University of Munich, Munich, Germany. <sup>2</sup>Medical Department 2 (Hematology/Oncology and Infectious Diseases), Center for Internal Medicine, Goethe University Frankfurt, University Hospital, Frankfurt am Main, Germany. <sup>3</sup>German Cancer Consortium (DKTK), Partner Site Frankfurt/Mainz and German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>4</sup>Department I for Internal Medicine, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany. <sup>5</sup>German Centre for Infection Research (DZIF), Partner Site Bonn-Cologne, Cologne, Germany. <sup>6</sup>Division of Dermatology and Venereology, Department of Medicine Solna, Karolinska Institutet, Stockholm, Sweden. ✉email: alexander.zink@tum.de

disease awareness<sup>28,29</sup>. However, individuals translate health-related information needs into web search requests with the intention to identify and access respective information<sup>30,31</sup>. Thus, web searches reflect the engagement between users and information sources and in this sense can be seen as “individual proxies for public disease awareness”<sup>32,33</sup>. Individual motivations of searches (mere interest-driven vs. problem-driven) or the quality of consecutively accessed information resources cannot be derived solely from SV.

In this sense, we consider SV as an economical and accessible data source to approximate public disease awareness that may help to plan prevention measures effectively.

However, to the best of our knowledge, the relationship between the incidence of various cancers and SV as proxy of public awareness has not yet been studied in Germany. The aim of this study was to characterize the spatiotemporal relationship between incidence and SV in Germany for 8 cancer entities: breast, prostate, colorectal, lung, cervical, and testicular cancer, melanoma, and brain tumors. Incidence data were provided by the German Center for Cancer Registry Data (German: *Zentrum für Krebsregisterdaten*, ZfKD), and SV was retrieved from Google’s Ads Keyword Planner. Specifically, we examined the incidence and SV for each of these cancers, assessed temporal and regional patterns, and investigated the association between cancer-specific incidence and SV. To identify themes of high public interest, we explored the content of the entity-specific web searches.

## Methods

### Data

This retrospective, observational study combined monthly data on cancer incidence and SV from July 2018 to December 2019 for district-free cities in Germany. District-free cities (German: *Kreisfreie Städte*) are units where the city proper is identical to the area of the administrative unit. We restricted our analysis to district-free cities, as they were the smallest spatial unit for which both cancer incidence and SV data were available and could be linked. With respect to the negative association of rural living on health care utilization<sup>34</sup> and the heterogeneity of internet coverage in urban vs. rural areas<sup>35</sup>, limiting the analysis to district free cities aimed to avoid these potential confounders. Data from all district-free cities (N = 107) were investigated (Fig. 1). The total population of all district-free cities was 26.9 million inhabitants, representing about one-third of the German population<sup>36</sup>.

Based on previous findings that disease incidence, prevention campaigns, and disease-related survival could drive disease awareness<sup>16,18–20,24,37</sup>, we selected the above-mentioned cancer entities based on their respective incidence, survival, and prevention campaigns (high incidence: breast, prostate, colorectal, lung, and melanoma; low disease-related survival: lung cancer and brain tumor; prevention campaigns: breast, prostate, colorectal, cervical, and testicular cancer and melanoma).

### Cancer incidence data

Data on cancer incidence by district and month were provided on request by the ZfKD, at the Robert Koch Institute<sup>38</sup>. The ZfKD’s data collection procedure is detailed in<sup>39</sup>; details on the registry coverage can be found in<sup>40</sup>. Diagnoses were based on the International Classification of Diseases 10th revision (ICD-10, Supplementary Table S1). In addition, we extracted the month and year of diagnosis and the district area code of residence of each diagnosed person. We limited the data analysis to the period of July 2018 to December 2019 as congruent web search and registry data were only available for this period.

### Web search data

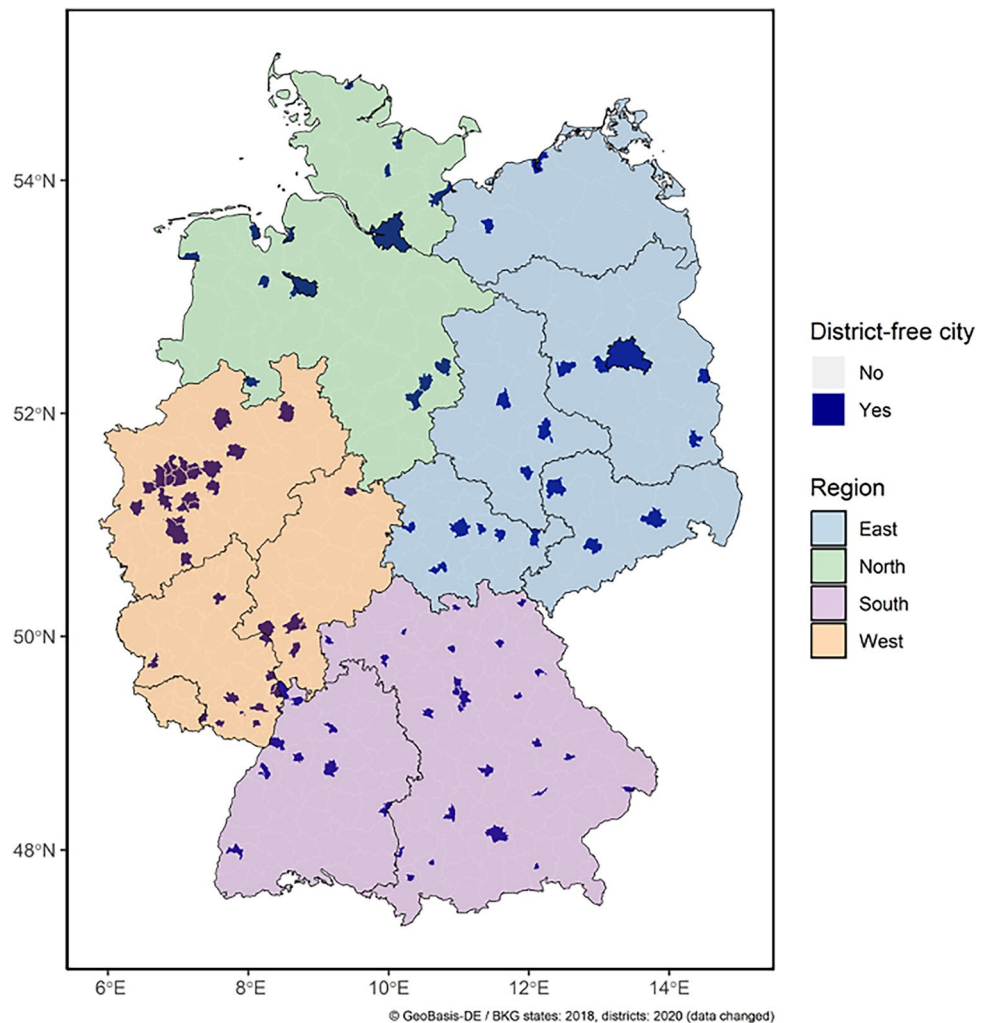
In Germany, Google is the most frequently used search engine with a market share above 90%<sup>41</sup>. SV data were, thus, based on queries to the Google search engine and its network partners and gathered via the Google Ads Keyword Planner<sup>17,42,43</sup>. For each search term entered, this tool provides a comprehensive list of keywords and phrases as well as the number of searches from a given geographic area in the past 48 months. To capture a broad range of search terms, we entered a medical and a lay term for each of the 8 above-mentioned malignancies (Supplementary Table S1). Data extraction was restricted to keywords and phrases in German and to web searches that could be traced back to one of the 107 district-free cities.

All identified keywords (k = 12,551) were manually reviewed. Keywords that were not directly linked to investigated diseases were excluded from further analysis (e.g., “antihormone therapy side effects”). For skin cancer-related web searches, we additionally excluded all keywords related to keratinocyte carcinoma (33.5%) and restricted the search terms to keywords related to melanoma.

### Data linkage

Cancer incidence and web search data were linked based on location and date, resulting in 18 monthly observations for each pair of cancer entity and district-free city. To ensure comparability between heterogeneously populated cities, we scaled incidence (Inc/100k) and SV (SV/100k) per 100,000 inhabitants. We extracted the population size of each city in 2019 from the database of the Federal Office for Building and Regional Planning (German: *Indikatoren und Karten zur Raum- und Stadtentwicklung*; INKAR)<sup>36</sup>. To gain insight into regional differences, the district-free cities were grouped by a common regional partition of Germany, roughly corresponding to cardinal directions. In total, there were 15 district-free cities in the North, 19 in the East, 34 in the South, and 39 in the West. For visualization of the selected district-free cities and federal states of Germany, publicly available geospatial data based on the German Federal Agency for Cartography and Geodesy were obtained from Esri Germany<sup>44</sup>.

To obtain access to the incidence data, the study protocol was handed in and approved by the scientific review board of the ZfKD. Institutional review board approval and informed consent were not required for web search data due to its non-disclosive nature and public availability.



**Figure 1.** Map of Germany with all  $N = 107$  district-free cities, with color-coded regions. In total, there were 15 district-free cities in the North, 19 in the East, 34 in the South, and 39 in the West. For visualization of the selected district-free cities and federal states of Germany, publicly available geospatial data based on the German Federal Agency for Cartography and Geodesy were obtained from Esri Germany.

### Statistical analysis

For each malignancy, we assessed both the time- and space-aggregated distributions of Inc/100k and SV/100k using median and interquartile range (IQR). To determine whether the incidence of malignancies is associated with SV, we computed Spearman's correlation coefficient ( $r_s$ ) between the time- and space-aggregated totals of absolute incidence and SV.

To examine the temporal patterns of cancer-specific incidence and SV, we visualized space-aggregated Inc/100k and SV/100k with line plots across months. We calculated normalized means and Gaussian 95% confidence intervals (CI). Depending on the underlying distribution, regional differences of time-aggregated cancer Inc/100k and SV/100k were assessed with ANOVA, Welch-ANOVA, or Kruskal–Wallis tests. To test for differences between regions, we performed Tukey, Games–Howell, or Dunn's post-hoc tests ( $P_{post-hoc}$ ) in follow-up analyses.  $P$  values of all post-hoc tests were corrected with the Bonferroni method. Before testing for regional differences using the respective methods, the assumption of approximate normality was assessed graphically and the homogeneity of variance was tested using Levene's test. To determine overall as well as region-specific associations between Inc/100k and SV/100k, repeated measures correlations ( $r_{rm}$ ) and corresponding 95% CIs were calculated for each cancer entity<sup>45</sup>. Additionally, the  $P$  values for the overall correlations were adjusted using Bonferroni correction and the region-specific correlations using Benjamini–Yekutieli correction, as their  $P$  values may be interdependent. For all analyses, two-sided tests were performed and the significance level was set to 0.05. Statistical analyses were performed using the statistical software R version 4.2.3 (R Core Team, 2021, Vienna, Austria).

The web search keywords were classified deductively and inductively and emerging topics were identified and discussed iteratively between the researchers (HW, DM, and SZ) until 15 final categories were determined: *diagnosis* (a diagnosis-indicating disease specification; e.g., “stage IV breast cancer”), *symptoms* (e.g., “weight

loss colon cancer”), *treatment* (e.g., “skin cancer treatment”), *prognosis* (e.g., “glioblastoma life expectancy”), *risk factors/triggers* (e.g., “lung cancer smoking”), *demographics* (e.g., “lung cancer women”), *comorbidity* (e.g., “Crohn’s disease colon cancer”), *consequences* (e.g., “impotence after prostate cancer”), *prevention* (e.g., “hpv pap smear cervical cancer”), *costs* (e.g., “surgery costs colon cancer”), *celebrities* (e.g., “Kylie Minogue breast cancer”), *general information* (e.g., “melanoma”), *media* (e.g., “brain tumor documentary”), *peer-community* (e.g., “lung cancer experience reports”), and *others* (e.g., “prostate cancer cycling”). Finally, one researcher (HW) resolved conflicting keyword categorizations by different coders to ensure that each search term was exclusively assigned to one category. The content of malignancy-related web searches was analyzed descriptively by calculating the percentage of time- and space-aggregated SV/100k for each category within a malignancy.

## Results

### Descriptive analysis of cancer incidence and web search volume

In total, 126,350 inhabitants of German district-free cities were diagnosed with cancer between July 2018 and December 2019 (Table 1). During the same period, a total of 21,116,930 malignancy-related web searches were recorded in these district-free cities, after excluding 1,256 irrelevant keywords out of a total of 12,759 cancer-related German keywords and phrases.

Between July 2018 and December 2019, breast, prostate, lung, and colorectal cancer had the greatest median Inc/100k. Median SV/100k were highest for breast cancer, lung cancer, and melanoma. On the other end, we observed the lowest incidence rates and SV for cervical and testicular cancer (Table 1). Cancer incidence was strongly associated with SV as reflected by a Spearman’s correlation coefficient of  $r_s = 0.88$  ( $P = 0.007$ ).

### Temporal and spatial patterns of cancer incidence and search volume

For all malignancies, normalized mean Inc/100k and SV/100k showed similar patterns across months with marked decreases in June, August, September, and December (Fig. 2). However, the normalized mean SV/100k for melanoma increased substantially in the summer months.

Regional differences in Inc/100k were most pronounced for breast cancer, prostate cancer, and melanoma (Fig. 3), with the lowest incidence rates in the East compared to other regions (breast:  $P_{post-hoc} < 0.001$ ; prostate:  $0.03 \leq P_{post-hoc} \leq 0.04$ ; melanoma:  $0.003 \leq P_{post-hoc} \leq 0.03$ ). We also found regional differences in the incidence of lung cancer ( $P_{post-hoc} < 0.001$ ), though not consistently across all region pairs. A lower cancer incidence was recorded in Eastern Germany for 5 of 8 malignancies.

Differences in SV/100k between German regions were observed for the majority of malignancies. In particular, we found differences between Southern Germany and the other regions with a higher SV/100k for brain tumors ( $P_{post-hoc} \leq 0.01$ ), breast cancer ( $P_{post-hoc} \leq 0.03$ ), prostate cancer ( $P_{post-hoc} \leq 0.02$ ), and melanoma ( $P_{post-hoc} \leq 0.02$ ). SV/100k for lung cancer differed between the South and East ( $P_{post-hoc} = 0.002$ ) and between the South and West ( $P_{post-hoc} = 0.02$ ).

Comparing the temporal patterns per region (Supplementary Fig. S2), we observed similar trajectories of normalized mean Inc/100k and normalized mean SV/100k for most malignancies. For instance, we detected a similar pattern of cancer incidence and normalized search volume for prostate cancer in the West and South. Across all regions, we found seasonal variations with cancer Inc/100k and SV/100k decreasing in June, August, September, and December. Furthermore, search volume for lung cancer peaked in November 2018 in all regions.

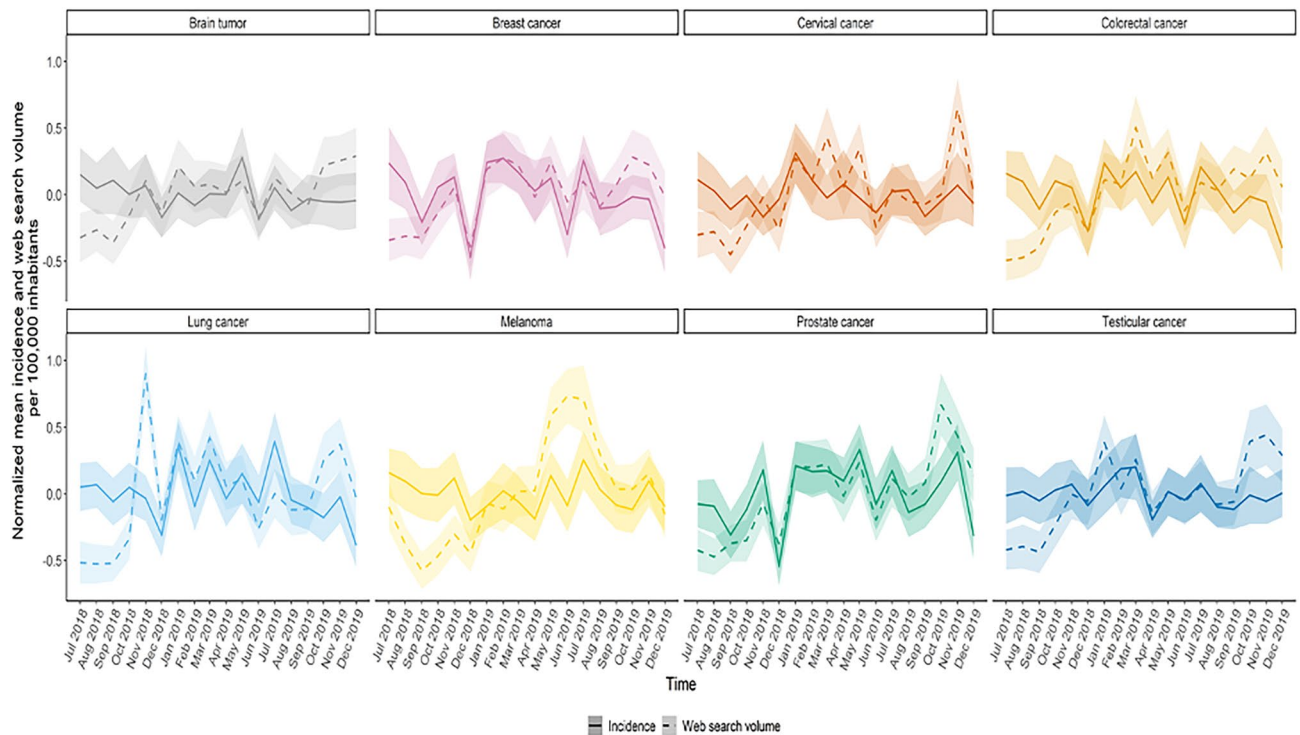
### Association analysis

For all 8 entities, greater SV/100k was associated with higher Inc/100k (Table 2). Only for colorectal cancer was this positive correlation not significant ( $r_{rm} = 0.037$ , 95% CI:  $[-0.009, 0.082]$ ). The strongest correlations between Inc/100k and SV/100k were observed for testicular ( $r_{rm} = 0.208$ ,  $[0.163, 0.251]$ ), prostate ( $r_{rm} = 0.196$ ,  $[0.152, 0.240]$ ), and breast cancer ( $r_{rm} = 0.109$ ,  $[0.064, 0.155]$ ).

Malignancy	Total incidence, n (rank)	Incidence per 100,000 inhabitants, median [IQR]	Total search volume, n (rank)	Search volume per 100,000 inhabitants, median [IQR]
Brain tumor	2891 (6)	0.40 [0.00; 0.92]	2,743,900 (5)	635.90 [525.13; 822.60]
Breast cancer	32,511 (1)	6.80 [5.06; 8.53]	4,314,110 (1)	1,054.90 [891.26; 1,352.41]
Cervical cancer	2115 (7)	0.00 [0.00; 0.65]	1,278,560 (7)	306.80 [246.94; 404.55]
Colorectal cancer	24,906 (4)	5.30 [3.94; 7.00]	2,499,140 (6)	607.30 [496.96; 759.83]
Lung cancer	27,121 (2)	5.40 [3.71; 7.48]	3,494,020 (2)	852.10 [707.91; 1,054.72]
Melanoma	9101 (5)	1.90 [1.00; 2.97]	3,280,220 (3)	763.20 [612.82; 974.52]
Prostate cancer	25,889 (3)	5.80 [4.24; 7.52]	2,770,460 (4)	691.80 [554.24; 898.67]
Testicular cancer	1816 (8)	0.0 [0.00; 0.60]	736,520 (8)	172.60 [135.53; 231.38]
Total	126,350		21,116,930	

**Table 1.** Malignancy-specific incidence and web search volume in total and time- and space-aggregated reported with median per 100,000 inhabitants including interquartile range between July 2018 and December 2019. The ranks of each entity’s total incidence and search volume are displayed in round brackets, from highest (1) to lowest (8). IQR interquartile range.





**Figure 2.** Malignancy-specific normalized mean and 95% confidence interval of space-aggregated incidence and search volume per 100,000 inhabitants across the months between July 2018 and December 2019.

The associational strength between Inc/100k and SV/100k varied across German regions (Table 2). For most cancer entities (6 of 8), the North or East showed stronger correlations than the other regions. The direction of these correlations was fairly consistent: 30 of 32 region-entity pairs had a positive correlation between Inc/100k and SV/100k. Yet for about 72% of the region-entity combinations, correlation between Inc/100k and SV/100k were not significant.

### Search content analysis

The content of web searches differed across malignancies (Fig. 4). For the majority of cancer entities, the most frequently queried keywords were related to cancer *diagnosis*, *symptoms*, and *general* information. These search categories were among the top 3 for 7 (*diagnosis*, *symptoms*) and 6 (*general* information) of the 8 entities, respectively. For brain tumors and lung cancer, searches related to *prognosis* were more frequent than in the other entities. The percentage of entity-specific web searches related to *prevention* was highest for cervical and colorectal cancer. For breast and prostate cancer, searches mainly focused on *treatment* options.

## Discussion

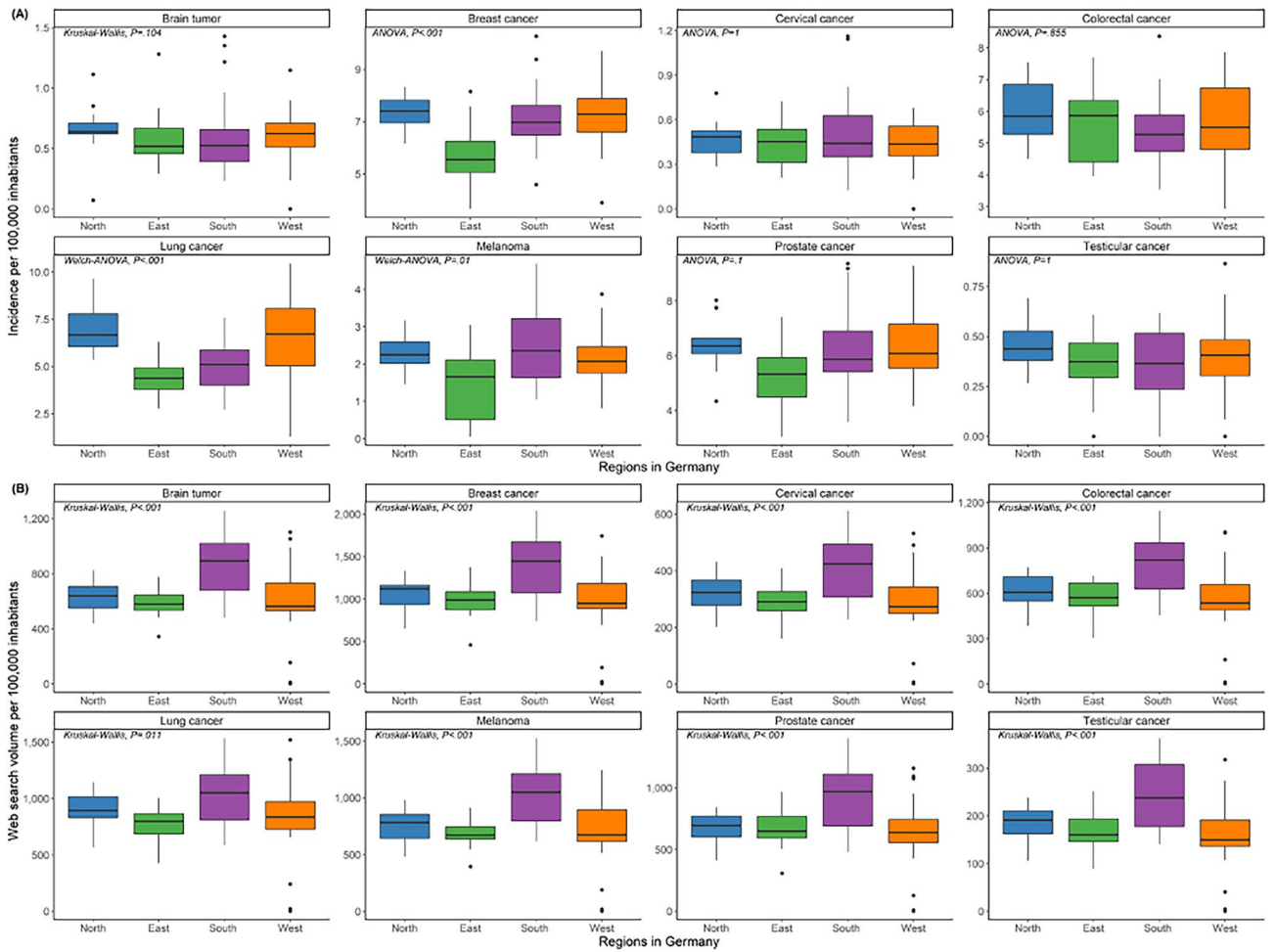
### Main results

In this study, we investigated spatiotemporal relationships between entity-associated SV and cancer incidence for 8 malignancies in Germany. Across all 8 entities, we observed that greater incidence rates were associated with higher SV. Specifically, more highly incident cancer entities, such as breast and lung cancer, were more frequently queried than rarer entities, such as testicular and cervical cancer.

The trajectory of incidence and SV followed a seasonal pattern: incidence and SV declined in the summer (July and August 2019) and winter months (December 2018, December 2019). A study in Sweden observed a similar pattern and attributed reduced incidence rates to lower detection rates during holiday seasons<sup>46</sup>. While we cannot draw causal inferences from our data, we may hypothesize that reduced incidences in Germany may be due to restricted opening times of doctors' offices and less staffing during the holiday season.

In line with previous studies, we found higher incidence rates and SV for melanoma during the summer<sup>11,42,43</sup>. We expected to observe this pattern as solar ultraviolet radiation is a major risk factor for melanoma and awareness campaigns for skin cancer are particularly active in the summer<sup>42</sup>. Additionally, we note that elevated SV coincided with prevention campaigns (e.g., breast cancer awareness month in October, lung cancer awareness month in November, world cancer day in February)<sup>21–23</sup> and media reports. For example, a spike in lung cancer SV could be observed in November 2018, the same month as a highly publicized death of a German celebrity to lung cancer<sup>47</sup>.

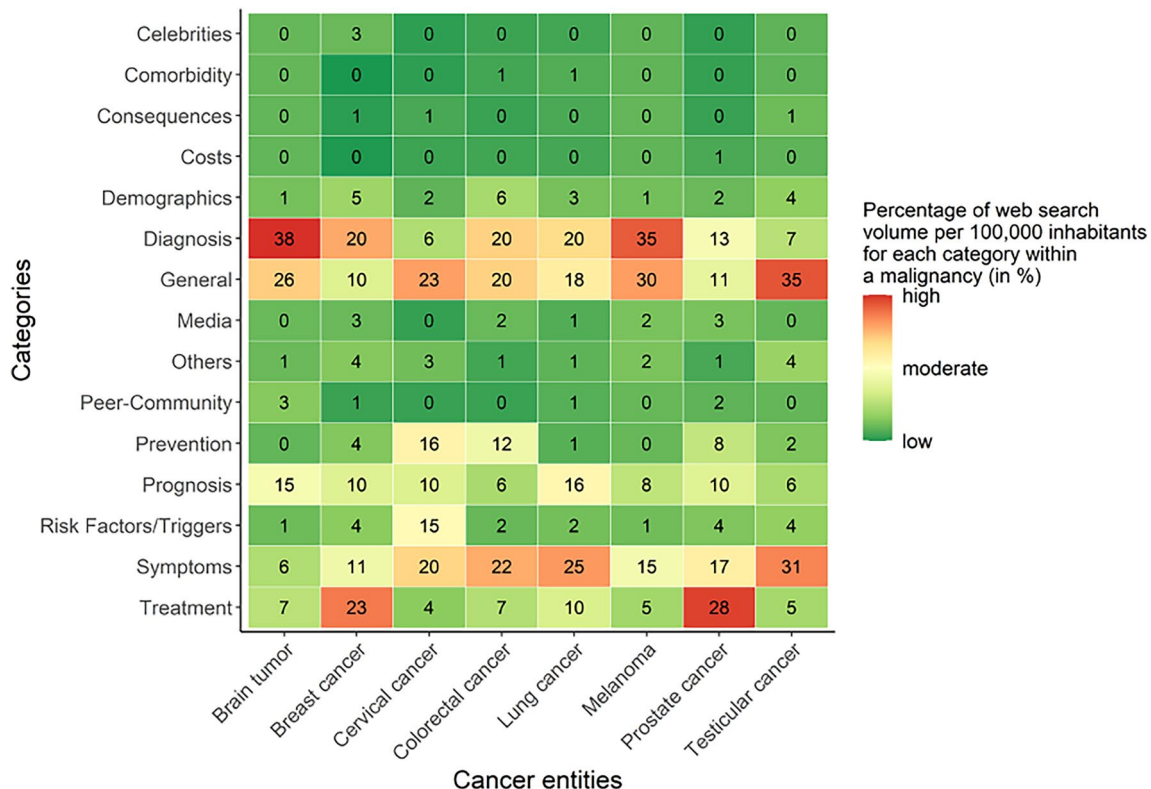
Across German regions, a lower cancer incidence has been indicated in Eastern Germany for most malignancies. This is partly in line with a previous study that reported a lower cancer incidence in women living in Eastern Germany compared to women in Western Germany, though the opposite trend was found in men<sup>48</sup>. Furthermore, Vogt et al. found higher screening rates for various cancers in districts in Eastern Germany than districts in the



**Figure 3.** Malignancy-specific boxplots of cancer incidence (A) and search volume (B) per 100,000 inhabitants for each of the regions: North, East, South, and West Germany. The respective statistical methods and *P* values to assess differences between regions are reported.

Malignancy Region	Repeated Measures Correlation, $r_{rm}$ [95% CI], <i>P</i> value				
	Overall <sup>a</sup>	North <sup>b</sup>	East <sup>b</sup>	South <sup>b</sup>	West <sup>b</sup>
Brain tumor	0.095 [0.049, 0.140], <i>P</i> < 0.001	0.093 [-0.029, 0.214], <i>P</i> = 0.735	0.131 [0.023, 0.237], <i>P</i> = 0.157	0.048 [-0.034, 0.129], <i>P</i> = 1	0.167 [0.092, 0.240], <i>P</i> < 0.001
Breast cancer	0.109 [0.064, 0.155], <i>P</i> < 0.001	0.189 [0.068, 0.305], <i>P</i> = 0.034	0.108 [-0.001, 0.214], <i>P</i> = 0.339	0.093 [0.011, 0.173], <i>P</i> = 0.207	0.111 [0.035, 0.186], <i>P</i> = 0.054
Cervical cancer	0.087 [0.042, 0.133], <i>P</i> = 0.002	0.129 [0.007, 0.248], <i>P</i> = 0.298	0.097 [-0.012, 0.204], <i>P</i> = 0.475	0.084 [0.002, 0.164], <i>P</i> = 0.314	0.064 [-0.012, 0.139], <i>P</i> = 0.566
Colorectal cancer	0.037 [-0.009, 0.082], <i>P</i> = 0.949	0.074 [-0.049, 0.195], <i>P</i> = 1	0.078 [-0.031, 0.185], <i>P</i> = 0.840	0.014 [-0.068, 0.095], <i>P</i> = 1	0.036 [-0.041, 0.111], <i>P</i> = 1
Lung cancer	0.082 [0.037, 0.128], <i>P</i> = 0.004	0.168 [0.047, 0.285], <i>P</i> = 0.075	0.199 [0.092, 0.301], <i>P</i> = 0.005	-0.002 [-0.083, 0.080], <i>P</i> = 1	0.101 [0.025, 0.176], <i>P</i> = 0.092
Melanoma	0.070 [0.024, 0.116], <i>P</i> = 0.021	-0.041 [-0.162, 0.082], <i>P</i> = 1	0.132 [0.023, 0.238], <i>P</i> = 0.157	0.081 [0, 0.162], <i>P</i> = 0.339	0.071 [-0.005, 0.146], <i>P</i> = 0.416
Prostate cancer	0.196 [0.152, 0.240], <i>P</i> < 0.001	0.173 [0.051, 0.289], <i>P</i> = 0.066	0.026 [-0.083, 0.135], <i>P</i> = 1	0.206 [0.126, 0.283], <i>P</i> < 0.001	0.273 [0.201, 0.342], <i>P</i> < 0.001
Testicular cancer	0.208 [0.163, 0.251], <i>P</i> < 0.001	0.273 [0.156, 0.383], <i>P</i> < 0.001	0.297 [0.194, 0.393], <i>P</i> < 0.001	0.172 [0.092, 0.250], <i>P</i> = 0.001	0.172 [0.097, 0.245], <i>P</i> < 0.001

**Table 2.** Repeated measures correlation ( $r_{rm}$ ) between cancer incidence and search volume per 100,000 inhabitants, including 95% confidence interval (CI) and *P* values after correction for multiple testing overall and for each German region, cardinally classified into North, East, South, and West Germany. <sup>a</sup>Correction for multiple testing via Bonferroni correction. <sup>b</sup>Correction for multiple testing via Benjamini–Yekutieli correction.



**Figure 4.** Percentage of malignancy-specific search volumes per 100,000 inhabitants for each search category. Keywords were classified deductively and inductively into 15 categories: diagnosis (a diagnosis-indicating disease specification; e.g., “stage IV breast cancer”), symptoms (e.g., “weight loss colon cancer”), treatment (e.g., “skin cancer treatment”), prognosis (e.g., “glioblastoma life expectancy”), risk factors/triggers (e.g., “lung cancer smoking”), demographics (e.g., “lung cancer women”), comorbidity (e.g., “Crohn’s disease colon cancer”), consequences (e.g., “impotence after prostate cancer”), prevention (e.g., “hpv pap smear cervical cancer”), costs (e.g., “surgery costs colon cancer”), celebrities (e.g., “Kylie Minogue breast cancer”), general information (e.g., “melanoma”), media (e.g., “brain tumor documentary”), peer-community (e.g., “lung cancer experience reports”), and others (e.g., “prostate cancer cycling”). Colors are only interpretable within one cancer entity. Percentages may not sum up to 100 due to rounding.

West<sup>49</sup>, which may result in lower incidence rates due to early detection of precancerous lesions<sup>50</sup>. We also found regional differences in web searches, with higher SV for all malignancies in Southern Germany. Compared to the other regions, Southern Germany has the youngest inhabitants<sup>36</sup>. Younger individuals have been found to search more frequently for health information on the internet than older individuals<sup>8,51</sup>. Thus, differences in SV across regions may be at least partially attributable to demographic differences. Further research is required to study how regional differences in demographic characteristics, such as sex, age, and socioeconomic status, as well as the accessibility and distribution of medical care facilities, are linked to SV and cancer incidence.

We observed statistically significant small to moderate associations between cancer Inc/100k and SV/100k for all malignancies except colorectal cancer. For these entities, increases in SV/100k were associated with rising Inc/100k. Importantly, the strength of this correlation differed across regions. For breast and cervical cancer, we found the strongest correlation between cancer incidence and SV in Northern Germany, whereas the association for melanoma, colorectal, lung, and testicular cancer were strongest in Eastern Germany. While we observe a link between cancer incidence and SV, other factors such as access to close-by health care facilities, health promotion (media) campaigns etc. certainly play a major role in steering public awareness. Further research is needed to investigate the causes of regional differences and the interplay of online health seeking behavior with health information promoting factors.

Web search content differed across cancer entities. Comparatively, web searches related to brain tumors and lung cancer focused relatively more often on disease prognosis. This may be due to their low disease-related survival probabilities. On the other hand, web searches related to entities with multiple available therapy options—such as breast and prostate cancer—particularly targeted information about treatments<sup>10–12</sup>. In addition, web searches related to cervical and colorectal cancer also focused on preventive measures, such as screening<sup>52</sup>.

### Limitations

Our analysis is based on observational, aggregated data at a population level. Thus, reliable statements about causal relationships between cancer incidence and SV cannot be derived.

Cancer incidence data were provided by the ZfKD, which brings together cancer registry data at the national level and can be regarded as a source of reliable, standardized data. SV in contrast is a type of digital trace data that can only serve as a surrogate measure for disease awareness.

Our results may not generalize to all German regions and cancer entities:

We used data from Germany's district-free cities ( $N = 107$ ) and did not include smaller urban and rural areas. Thus, our findings may not be representative of the entire German population. We also restricted our analysis to 8 cancer entities, which were selected based on incidence, prevention measures (i.e., campaigns), and survival. Across regions and entities, we observed differing strengths of association between web searches and cancer incidence. The observed patterns may not extend to other cancer entities.

Further, SV was based on German search terms only. Thus, we cannot infer information on the search behavior of non-German speakers. Additionally, our inferences rely on German search results being representative of the search behavior of German-speaking inhabitants of the cities investigated.

Due to the restricted overlap of available SV and cancer incidence data, we could only analyze time series data with 18 monthly time steps. This precluded the application of time series regression and the robustness of seasonal effects analysis. Rigorous testing of the observed declines in cancer incidence during the German holiday months and the seasonal elevation of melanoma awareness in summer as reflected in SV are still needed.

Previous research finds that young, female, and highly educated adults are more likely to use the internet to search for health information than older, male, and less educated individuals<sup>16,30,51,53</sup>. We cannot infer the characteristics of the individuals, on whom the web search data in this study was based. However, we speculate that information searching behavior of older individuals, who are typically at a higher risk for cancer, is under-represented in this study.

Epidemiological studies have shown that lung and colorectal cancer incidence and mortality increase with higher deprivation of living area<sup>54,55</sup>. Area deprivation also is negatively associated with access to health care<sup>34</sup>. Thus, living area deprivation could be a unobserved confounder in our study, that future research should address.

### Comparison with prior work

Our results contribute to an existing body of similar research from other geographical contexts. While studies in the US<sup>19,37</sup> and China<sup>16</sup> have observed strongly positive correlations between cancer incidence and SV, we only found low to moderate associations, comparable to those reported by Phillips et al.<sup>18</sup>. This may be attributable to our methodological approach of narrowing the geographical scope to district-free cities instead of the state level. We examined the association between SV and cancer incidence across months with repeated measures correlations. Interestingly, after adjusting for the measurement period, Phillips et al.<sup>18</sup> also reported substantially lower correlation coefficients.

Online information-seeking behavior on multiple cancer entities has been studied through the lens of Google data in and across various countries—in particular in the US<sup>18,20,37</sup>, but also in China<sup>16</sup> and Canada<sup>56</sup>. Prior research on disease-related and web-based information-seeking predominantly used Google Trends<sup>18–20,24,25,37,56,57</sup> instead of Google Ads Keyword Planner<sup>42,43</sup>. We considered both data sources but decided on the latter, as it provides more fine-grained information, including geographical units and absolute counts instead of relative frequency abstractions. Moreover, Google Ads Keyword Planner allows for the analysis of keywords and phrases related to search terms, which provides more detailed insight into the public's interests, unmet needs, and awareness of different types of cancer<sup>17</sup>. Thus, to the best of our knowledge, this study is not only the first to examine the association between SV and multiple cancer entities, but also the first to assess the relationship at this level of geographical detail.

### Conclusions

We report moderate but consistent positive associations between cancer incidence and SV for a set of cancer entities in German district-free cities. Examining the relationship between SV—as a proxy of public awareness—and incidence—as one of its key drivers—reveals regions with higher disparity between cancer incidence and SV. Such disparities could signify socially-deprived areas with unmet information needs<sup>24</sup> and could be of use for planning future prevention measures.

Our investigation is a first step toward using web search information to understand German public awareness about cancer. Further research is needed to describe more precisely the relationship between public awareness and web search behavior. This will likely require the development and empirical validation of models that describe the interplay of social dynamics, perception, and group and individual actions.

### Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Received: 23 January 2024; Accepted: 21 April 2024

Published online: 03 May 2024

### References

1. Statistical Office of the European Union. Causes of death statistics (2023). [https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Causes\\_of\\_death\\_statistics](https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Causes_of_death_statistics) [accessed 16-06-2023].
2. Erdmann, F., Spix, C., Katalinic, A., Christ, M., Folkerts, J., Hansmann, J., Kranzhöfer, K., Kunz, B., Manegold, K., Penzkofer, A., Treml, K., Vollmer, G., Weg-Remers, S., Barnes, B., Buttman-Schweiger, N., Dahm, S., Fiebig, J., Franke, M., Gurung-Schönfeld, I., Haberland, J., Imhoff, M., Kraywinkel, K., Starker, A., von Berenberg-Gossler, P. & Wienecke, A. Krebs in Deutschland für 2017/2018 (2021).



3. German Cancer Research Center. Erstmals für Deutschland ermittelt: Vermeidbare Risikofaktoren verursachen 37 Prozent aller Krebsfälle (2018). <https://www.dkfz.de/de/presse/pressemitteilungen/2018/dkfz-pm-18-48-Vermeidbare-Risikofaktoren-verursachen-37-Prozent-aller-Krebsfaelle.php> [accessed 15-06-2023].
4. Mons, U., Gredner, T., Behrens, G., Stock, C. & Brenner, H. Cancers due to smoking and high alcohol consumption. *Dtsch. Arztebl. Int.* **115**(35–36), 571–577 (2018).
5. Behrens, G. *et al.* Cancers due to excess weight, low physical activity, and unhealthy diet. *Dtsch. Arztebl. Int.* **115**(35–36), 578–585 (2018).
6. Gredner, T., Behrens, G., Stock, C., Brenner, H. & Mons, U. Cancers due to infection and selected environmental factors. *Dtsch. Arztebl. Int.* **115**(35–36), 586–593 (2018).
7. Bachl, M. Online health information seeking in Europe: Do digital divides persist?. *SCM Stud. Commun. Media* **5**(4), 427–453. <https://doi.org/10.5771/2192-4007-2016-4-427> (2016).
8. Link, E. & Baumann, E. Nutzung von Gesundheitsinformationen im Internet: Personenbezogene und motivationale Einflussfaktoren. [Use of health information on the internet: Personal and motivational influencing factors]. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* **63**(6), 681–689 (2020).
9. Perrault, E. K., Hildenbrand, G. M., McCulloch, S. P., Schmitz, K. J. & Lambert, N. J. Online information seeking behaviors of breast cancer patients before and after diagnosis: From website discovery to improving website information. *Cancer Treat. Res. Commun.* **23**, 100176 (2020).
10. Eysenbach, G. The impact of the Internet on cancer outcomes. *CA Cancer J. Clin.* **53**(6), 356–371 (2003).
11. Zhang, H. *et al.* Research interest and public interest in melanoma: A bibliometric and Google trends analysis. *Front. Oncol.* **11**, 629687 (2021).
12. Nanda, J. K., Hay, J. L. & Marchetti, M. A. Analysis of keywords used in internet searches for melanoma information: Observational study. *JMIR Dermatol.* **4**(1), e25720 (2021).
13. Tan, S. Y., Tang, M. S. S., Ong, C.-A. J., Tan, V. K. M. & Shannon, N. B. Impact of COVID-19 on public interest in breast cancer screening and related symptoms: Google trends analysis. *JMIR Cancer* **9**, e39105 (2023).
14. Finney Rutten, L. J. *et al.* Cancer-related information seeking among cancer survivors: Trends over a decade (2003–2013). *J. Cancer Educ.* **31**(2), 348–357 (2016).
15. Castleton, K. *et al.* A survey of Internet utilization among patients with cancer. *Support. Care Cancer* **19**(8), 1183–1190 (2011).
16. Xu, C. *et al.* Association between cancer incidence and mortality in web-based data in China: Infodemiology study. *J. Med. Internet Res.* **21**(1), e10677 (2019).
17. Ziehfrennd, S., Tizek, L. & Zink, A. Websearch-Daten als Gesundheitsdaten?: Geografische Unterschiede, zeitliche Trends und Interessenschwerpunkte von Internetsuchmaschinenanfragen in Deutschland. [Web search data as health data?: Geographic differences, temporal trends, and topics of interest from internet search engine analyses in Germany]. *Hautarzt* **73**(1), 53–60 (2022).
18. Phillips, C. A. *et al.* Relationship between state-level Google online search volume and cancer incidence in the United States: Retrospective study. *J. Med. Internet Res.* **20**(1), e6 (2018).
19. Xu, C. *et al.* Detecting lung cancer trends by leveraging real-world and internet-based data: Infodemiology study. *J. Med. Internet Res.* **22**(3), e16184 (2020).
20. Xu, C. *et al.* Leveraging Internet search data to improve the prediction and prevention of noncommunicable diseases: Retrospective observational study. *J. Med. Internet Res.* **22**, e18998 (2020).
21. Patel, M. S. *et al.* Success of prostate and testicular cancer awareness campaigns compared to breast cancer awareness month according to internet search volumes: A Google trends analysis. *Urology* **139**, 64–70 (2020).
22. Nishimura, Y. & Acoba, J. D. Impact of breast cancer awareness month on public interest in the united states between 2012 and 2021: A Google trends analysis. *Cancers (Basel)* **14**(10), 2534 (2022).
23. Glynn, R. W., Kelly, J. C., Coffey, N., Sweeney, K. J. & Kerin, M. J. The effect of breast cancer awareness month on internet search activity—A comparison with awareness campaigns for lung and prostate cancer. *BMC Cancer* **11**, 442 (2011).
24. Mondia, M. W. L., Espiritu, A. I. & Jamora, R. D. G. Brain tumor infodemiology: Worldwide online health-seeking behavior using Google trends and Wikipedia pageviews. *Front. Oncol.* **12**, 855534 (2022).
25. Gianfredi, V., Nucci, D., Nardi, M., Santangelo, O. E. & Provenzano, S. Using Google trends and Wikipedia to investigate the global public's interest in the pancreatic cancer diagnosis of a celebrity. *Int. J. Environ. Res. Public Health* **20**(3), 2106 (2023).
26. Sajjadi, N. B. *et al.* Public interest and behavior change in the United States regarding colorectal cancer following the death of Chadwick Boseman: Infodemiology investigation of internet search trends nationally and in at-risk areas. *JMIR Infodemiol.* **1**(1), e29387 (2021).
27. Lazer, D. *et al.* Meaningful measures of human society in the twenty-first century. *Nature* **595**(7866), 189–196 (2021).
28. Baiardini, I. *et al.* Disease awareness in patients with COPD: Measurement and extent. *Int. J. Chron. Obstruct. Pulmon. Dis.* **14**, 1–11 (2019).
29. Corli, O. *et al.* Illness awareness in terminal cancer patients: An Italian study. *Palliat. Med.* **23**(4), 354–359 (2009).
30. Bachl, M., Link, E., Mangold, F. & Stier, S. Search engine use for health-related purposes: Behavioral data on online health information-seeking in Germany. *Health Commun.* <https://doi.org/10.1080/10410236.2024.2309810> (2024).
31. Zhang, Y., Sun, Y. & Kim, Y. The influence of individual differences on consumer's selection of online sources for health information. *Comput. Hum. Behav.* **67**, 303–312. <https://doi.org/10.1016/j.chb.2016.11.008> (2017).
32. Silangcruz, K. *et al.* Impact of the world inflammatory bowel disease day and Crohn's and colitis awareness week on population interest between 2016 and 2020: Google trends analysis. *JMIR Infodemiol.* **1**(1), e32856 (2021).
33. Eysenbach, G. Infodemiology and infoveillance: Framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. *J. Med. Internet Res.* **11**(1), e11 (2009).
34. Greiner, G. G., Schwettmann, L., Goebel, J. & Maier, W. Primary care in Germany: Access and utilisation—a cross-sectional study with data from the German Socio-Economic Panel (SOEP). *BMJ Open* **8**(10), e021036 (2018).
35. Bundesnetzagentur. Breitbandatlas | Gigabit-Grundbuch. <https://gigabitgrundbuch.bund.de> [accessed 20-03-2024].
36. Bundesinstitut für Bau-, Stadt- und Raumforschung (BBSR) im Bundesamt für Bauwesen und Raumordnung. Indikatoren und Karten zur Raum- und Stadtentwicklung. INKAR. Ausgabe 2021. Bonn 2021. <https://www.inkar.de/> [accessed 18-11-2022].
37. Wehner, M. R., Nead, K. T. & Linos, E. Correlation among cancer incidence and mortality rates and internet searches in the United States. *JAMA Dermatol.* **153**(9), 911–914 (2017).
38. Zentrum für Krebsregisterdaten (ZfKD) im Robert Koch-Institut. Datensatz des ZfKD auf Basis der Epidemiologischen Landeskrebsregisterdaten, verfügbare Diagnosejahre bis 2019 (2022). Version: Epi2021\_1. <https://doi.org/10.18444/5.03.01.0005.0017.0001>.
39. Zentrum für Krebsregisterdaten (ZfKD) im Robert Koch-Institut. Methoden (2022). [https://www.krebsdaten.de/Krebs/DE/Content/Methoden/methoden\\_node.html](https://www.krebsdaten.de/Krebs/DE/Content/Methoden/methoden_node.html) [accessed 24-03-2023].
40. Zentrum für Krebsregisterdaten (ZfKD) im Robert Koch-Institut. Vollzähligkeitsschätzung. [https://www.krebsdaten.de/Krebs/DE/Content/Methoden/Vollzaehligkeitsschaetzung/vollzaehligkeitsschaetzung\\_node.html](https://www.krebsdaten.de/Krebs/DE/Content/Methoden/Vollzaehligkeitsschaetzung/vollzaehligkeitsschaetzung_node.html) [accessed 20-03-2024].
41. StatCounter GlobalStats. Search Engine Market Share Germany: Jan–Dec 2019. <https://gs.statcounter.com/search-engine-market-share/all/germany/2019> [accessed 24-03-2023].
42. Seidl, S., Schuster, B., Rütth, M., Biedermann, T. & Zink, A. What do Germans want to know about skin cancer? A nationwide google search analysis from 2013 to 2017. *J. Med. Internet Res.* **20**(5), e10327 (2018).

43. Tizek, L. *et al.* Interest in skin cancer in urban populations: A retrospective analysis of Google search terms in nine large German cities. *Acta Derm. Venereol.* **99**(9), 797–804 (2019).
44. Esri Deutschland GmbH. Open Data Portal: Bundesländer 2018, Kreisgrenzen 2020. [GeoBasis-DE/BKG]. <https://opendata-esri.dech.hub.arcgis.com> [accessed 26-10-2023].
45. Bakdash, J. Z. & Marusich, L. R. Repeated measures correlation. *Front. Psychol.* **8**, 456 (2017).
46. Wikén, I., Andersson, T.M.-L. & Radkiewicz, C. Seasonal effects on cancer incidence and prognosis. *Acta Oncol.* **62**(2), 103–109 (2023).
47. TV-Auswanderer Jens Büchner mit 49 Jahren an Krebs gestorben. *Süddeutsche Zeitung*. November 18, 2018.
48. Lampert, T., Müters, S., Kuntz, B., Dahm, S. & Nowossadeck, E. 30 years after the fall of the Berlin Wall: Regional health differences in Germany. *J. Health Monit.* **4**(Suppl 2), 2–23 (2019).
49. Vogt, V., Siegel, M. & Sundmacher, L. Examining regional variation in the use of cancer screening in Germany. *Soc. Sci. Med.* **110**, 74–80 (2014).
50. Ma, Z.-Q. & Richardson, L. C. Cancer screening prevalence and associated factors among US adults. *Prev. Chronic Dis.* **19**, E22 (2022).
51. Jia, X., Pang, Y. & Liu, L. S. Online health information seeking behavior: A systematic review. *Healthcare (Basel)* **9**(12), 1740 (2021).
52. Sahin, A. N., Sahin, A. S., Schwenter, F. & Sebahang, H. YouTube videos as a source of information on colorectal cancer: What do our patients learn?. *J. Cancer Educ.* **34**(6), 1160–1166 (2019).
53. van Eenbergen, M. C. *et al.* Comparing survivors of cancer in population-based samples with those in online cancer communities: Cross-sectional questionnaire study. *JMIR Cancer* **8**(1), e19379 (2022).
54. Michalski, N. *et al.* German Index of Socioeconomic Deprivation (GISD): Revision, update and applications. *J. Health Monit.* **7**(Suppl 5), 2–23 (2022).
55. Kuznetsov, L., Maier, W., Hunger, M., Meyer, M. & Mielck, A. Associations between regional socioeconomic deprivation and cancer risk: Analysis of population-based Cancer Registry data from Bavaria, Germany. *Prev. Med.* **53**(4–5), 328–330 (2011).
56. Xu, R., Brown, P., Baxter, N. & Sawka, A. M. Online public interest in cancer during the COVID-19 pandemic. *JCO Clin. Cancer Inform.* **5**, 695–700 (2021).
57. Hagiya, H., Koyama, T. & Otsuka, F. Impact of the cervical cancer awareness months on public interest in Japan: A Google Trends analysis, 2012–2021. *Sci. Rep.* **12**(1), 15391 (2022).

## Acknowledgements

The authors thank Sebastian Sitaru for generating the web search data. We thank the Center for Cancer Registry Data (Zentrum für Krebsregisterdaten) for granting permission to utilize their data for this project. Additionally, we express our gratitude to the German Cancer Consortium's (DKTK) Clinical Data Science Group for their invaluable support.

## Author contributions

HW and DM contributed equally to this work. HW, DM, SZ, and AZ contributed to the initial research question and created the study design. HW, DM, SZ, FAUF planned and carried out the data generation and statistical analysis. HW, DM, SZ, FAUF, IE, JV, AZ contributed to the interpretation of the results. HW and DM took the lead in writing the manuscript. All authors provided critical feedback and helped shape the research, analysis and manuscript.

## Funding

Open Access funding enabled and organized by Projekt DEAL. This work was funded by the Technical University of Munich, TUM School of Medicine and Health, Munich, Germany and the German Cancer Consortium (DKTK).

## Competing interests

Hannah Wecker (HW), Stefanie Ziehfreund (SZ), and Ian Erhard (IE) have no conflicts of interest to declare. Daniel Maier (DM) received speaker honoraria from Free University Berlin and travel compensation from IQVIA. Fabienne A.U. Fox (FAUF) received travel compensation from IQVIA. Jörg Janne Vehreschild (JJV) has received personal fees from Merck / MSD, Gilead, Pfizer, Astellas Pharma, Basilea, German Centre for Infection Research (DZIF), University Hospital Freiburg/Congress and Communication, Academy for Infectious Medicine, University Manchester, German Society for Infectious Diseases (DGI), Ärztekammer Nordrhein, University Hospital Aachen, Back Bay Strategies, German Society for Internal Medicine (DGIM), Shionogi, Molecular Health, Netzwerk Universitätsmedizin, Janssen, NordForsk, Biontech, APOGEPHA and grants from Merck / MSD, Gilead, Pfizer, Astellas Pharma, Basilea, German Centre for Infection Research (DZIF), German Federal Ministry of Education and Research (BMBF), Deutsches Zentrum für Luft- und Raumfahrt (DLR), University of Bristol, Rigshospitalet Copenhagen, Network University Medicine. Alexander Zink (AZ) has been an advisor and/or received speaker's honoraria and/or received grants and/or participated in clinical trials of the following companies: AbbVie, Ammiral, Amgen, Beiersdorf Dermo Medical, Bencard Allergie, BMS, Celgene, Eli Lilly, GSK, Janssen Cilag, Leo Pharma, Miltenyi Biotec, Novartis, Pfizer, Sanofi-Aventis, Takeda Pharma, UCB Pharma.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-60267-4>.

**Correspondence** and requests for materials should be addressed to A.Z.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024