Acta Diabetologica (2019) 56:1149–1154
https://doi.org/10.1007/s00592-019-01350-5

ORIGINAL ARTICLE

Can internet search engine queries be used to diagnose diabetes? Analysis of archival search data

Irit Hochberg[1] · Deeb Daoud[1] · Naim Shehadeh[1,2] · Elad Yom-Tov[3]

Received: 4 March 2019 / Accepted: 16 April 2019 / Published online: 15 May 2019

Abstract

Aims Diabetes is often diagnosed late. This study aimed to assess the possibility for earlier detection of diabetes from search data, using predictive models trained on large-scale data.

Methods We extracted all English-language queries made by people in the USA to Bing during 1 year and identified queries containing symptoms of diabetes. We compared the ability of four different prediction models (linear regression, logistic regression, decision tree and random forest) to distinguish between users who stated that they were diagnosed with diabetes and users who did not refer to diabetes or diabetes drugs but queried about at least one of the symptoms.

Results We identified 11,050 "new diabetes users" who stated they had been diagnosed with diabetes and approximately 11.5 million "control users" who queried about symptoms without querying for terms related to diabetes. Both the logistic regression and the random forest models were able to distinguish between the populations with an area under curve of 0.92 which translates to a positive predictive value of 56% at a false-positive rate of 1%. The model could identify patients up to 240 days before they mentioned being diagnosed.

Conclusions Some undiagnosed diabetes patients can be detected accurately according to their symptom queries to a search engine. Such earlier diagnosis, especially in cases of type 1 diabetes, could be clinically meaningful. The ability of search engines to serve as a population-wide screening tool could potentially be improved using additional data provided by users.

Keywords Diabetes · Symptoms · Digital health · Internet

Introduction

Diabetes is highly prevalent, with a prevalence of 73 cases per 1000 persons globally and an incidence of 6.7 new cases diagnosed per 1000 persons per year [1]. Most patients are diagnosed by blood tests performed as part of screening or for other medical reasons, but almost half of diabetes patients reported in hindsight having had symptoms before diagnosis [2] and in almost a third of cases a symptom was recorded in the medical record prior to diagnosis [3]. One third to one half of people with diabetes are unaware they have the disease for several years, with many presenting with diabetes complications at time of diagnosis [4].

Utilization of novel digital advances has sparked a great interest in the diabetes medical community [5–7]. It is presumed that the role of digital information generated by people in conjunction with artificial intelligence in medical diagnosis and treatment will increase in the next few decades, and it is therefore interesting to assess the validity of using web searches to recognize various clinical situations. A vast majority of internet users report that they query for medical symptoms online and try to self-diagnose symptoms [8]. Studies have demonstrated that people ask about diseases at roughly their incidence and about their drugs at the rate that they are prescribed [9]. Queries on medical topics have been used to study a range of medical questions,

Managed by Antonio Secchi.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00592-019-01350-5) contains supplementary material, which is available to authorized users.

✉ Irit Hochberg
i_hochberg@rambam.health.gov.il

[1] Institute of Endocrinology, Diabetes and Metabolism, Rambam Health Care Campus, 8 Ha'Aliya Street, POB 9602, 31096 Haifa, Israel

[2] Bruce Rappaport Faculty of Medicine, Technion – Israel Institute of Technology, Haifa, Israel

[3] Microsoft Research, Herzliya, Israel

including the measurement of the effectiveness of childhood flu vaccines, discovery of adverse side events [9], and the detection of risk factors for disease [10].

Recent studies have shown that queries in search engines by internet users can be used to detect several solid tumors [11, 12] and Parkinson's disease [13]. Drawing on the fact that people usually ask about a medical condition they have after they are diagnosed [9], these studies identify the positive class as those people who mention that they have a particular medical condition, sometimes adding to the group people whose demographics and queries suggest that they are also patients. The time of diagnosis is assumed to be shortly prior to the first query about the condition, and there is evidence that it is indeed so (unpublished data). The negative class is taken from people who asked about similar topics or that identified themselves as the spouse of a patient [14]. Thus, these studies are based on self-admission of illness, without linking the (usually anonymous) search data to medical records.

Here we investigated whether a future diagnosis of diabetes can be predicted from earlier searches on medical symptoms. We also compared the prevalence of queries on symptoms prior to a new diabetes diagnosis to prevalence of queries on other medical conditions that are highly symptomatic. This allows us to roughly interpolate the rate of symptomatic diabetes patients.

Methods

We extracted all queries made to Bing in English by people in the USA between May 1, 2017, and April 30, 2018. For each user, we recorded an anonymized user name, time and date of the query, and query text. We identified relevant diabetes symptom queries by finding queries which contained one or more terms related to weight loss, urinary frequency, polyuria, tiredness, malaise, somnolence, slow healing of wounds, hunger, erectile dysfunction, thirst or blurry vision. The terms were expanded to 24 terms which include layman terms according to the dictionary in [9] (see Supplementary Information for list).

The population of users with a new diagnosis of diabetes was defined as those who made a query indicating that they have diabetes during the last 2 months of the data period (March and April 2018), and who otherwise did not mention diabetes or a drug used to treat diabetes in the first 10 months of the data period. Queries that indicated diabetes were those which contained the phrases "I was diagnosed" or "I have" and "diabetes", excluding phrases which indicated possibility ("do I have diabetes", "I think I have diabetes"). We have evidence from other medical situations that a first declaration on diagnosis of a disease is usually made in the days around the actual diagnosis (unpublished data).

The control population is comprised of people who did not refer to diabetes or to diabetes drugs during these 12 months and did, however, query about at least one of the above-mentioned relevant symptoms.

Each symptom-querying user was represented through the number of times they queried for each of the symptoms. That is, a vector containing the number of times each symptom was queried by this user.

We compared four prediction models generated to distinguish between the symptom-querying diabetes patients and the control population. The dependent variable was whether the user declared he or she was diagnosed with diabetes, and the independent variables were the number of times they queried about each symptom. The models compared were linear regression, logistic regression, decision trees, and random forest (with 50 trees).

This study was approved by the Ethics Committee of the Technion, Israel Institute of Technology.

Data were analyzed using MATLAB version 9.4 with the Statistics Toolbox version 11.3.

Results

In the 1-year period, we identified 11,050 users with a query indicating they have a diagnosis of diabetes and approximately 11.5 million control users who queried about symptoms that could be related to diabetes. (Terms for diabetes symptom queries are listed in Table S1 in the supplement.) Of the new diabetes diagnosis population, only 377 users (3.4%) made at least one symptom query in the year before diagnosis. The numbers of users in each group appear in Table S2.

We compared the ability of four prediction models (linear regression, logistic regression, decision trees, and random forest) to distinguish between the symptom-querying diabetes patients and the control population. We used the receiver operating curve (ROC) and the area under curve (AUC) of the models as the measure of model performance. AUC represents the ability of the model to predict diabetes diagnosis, where a perfect model would have an AUC of 1.0. All models were tenfold cross-validated [15].

Figure 1 shows the ROC curves and Table 1 the positive predictive value (PPV) for the studied population at two cutoff points. The most accurate models were the random forest, with an AUC of 0.93, and the logistic regression with an AUC of 0.92. (The difference between them was non-significant.) The high AUC shows the ability of the score to accurately identify people with a definite future diagnosis of diabetes. The difference in AUC between the logistic regression model (second-ranked model) and linear regression (the third-ranked model) was statistically significant, ($P = 0.03$, one-sided Wilcoxon test [16],

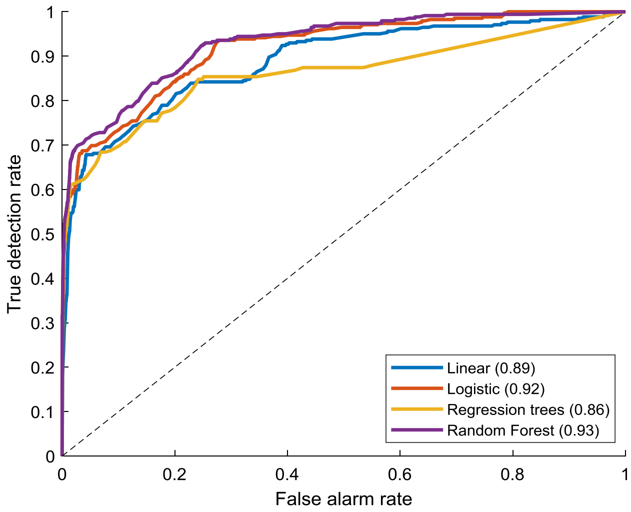

Fig. 1 Receiver operating curve (ROC) for distinguishing diabetes patients from the control population

Table 1 Performance comparison of the 4 predictive models

	AUC	PPV @0.01%	PPV @1%
Linear regression	0.89	16	46
Logistic regression	0.92	25	56
Decision tree	0.86	27	54
Random Forest	0.93	28	57

PPV positive predictive value

Since the random forest model and the logistic regression model have similar performance, and since the latter is interpretable, we report a detailed analysis of the parameters which were predictive in the logistic regression model (Table 2). Statistically significant symptom queries associated with a greater likelihood of being in the new diabetes diagnosis population were impotence, malaise, polyuria, and thirst. On the other hand, fatigue, hunger, somnolence, slow healing, and weight loss were associated with a lower likelihood of being in the new diabetes diagnosis population.

We performed sequential forward feature selection [15] to assess the contribution of each symptom to the prediction model. The results of this analysis are shown in Fig. 2, where the AUC is plotted from that achieved for the single best symptom (malaise) and as additional attributes are added, so that each added attribute increases the AUC by the greatest amount. Thus, queries about malaise, fatigue, and mentions of urine (in any context), and absence of queries about hunger and weight loss are the most indicative attributes.

Relative to the first date, a query mentioning diabetes diagnosis was made; we tested the likelihood of making queries for each of the positively associated symptoms with an increased likelihood of diabetes. Blurred vision, fatigue and hunger began to appear at high (> 1.5 times more likely) around 10 days prior to first date of diabetes query. At high likelihood, impotence appeared between 30 and 10 days prior to the first diabetes query. Interestingly, queries for weight loss began at least 2 months prior to the first diabetes query and maintain a level of approximately 50% more queries than expected by chance up until the first diabetes query.

To estimate how early diabetes could be predicted, we repeated the prediction process, after first removing queries made between the first diabetes query date and a varying number of days, from 5 to 240. The results for a logistic regression model are shown in Fig. 2. The AUC decreased for the longer time frames but was still 0.88 for 240 days.

Table 2 Model parameters for predicting diabetes diagnosis from queries on specific terms

Variable	Exp(coefficient) (SE)
Blurred vision	0.150 (0.037)
Fatigue	− 4.382 (0.077)
Hunger	− 3.459 (0.069)
Impotence	3.739 (0.220)
Malaise	0.677 (0.009)
Polyuria	2.396 (0.007)
Slow heal	− 0.485 (0.067)
Somnolence	− 1.784 (0.036)
Thirst	0.378 (0.024)
Urine	− 0.708 (0.012)
Weight loss	− 0.706 (0.012)

Positive exp(coefficient) indicates an increased likelihood of future diabetes diagnosis, and the magnitude of the slope indicates the relative importance of the symptom

All p values are smaller than 5×10^{-5}

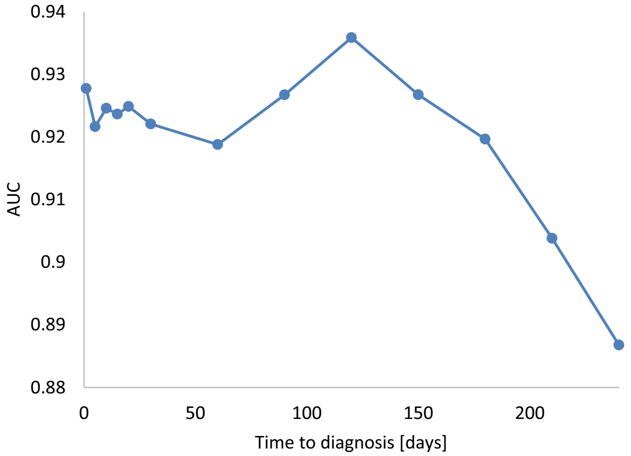

Fig. 2 Area under curve (AUC) as a function of the time between the last symptom query included in the model and the first query for diabetes

This indicates that in some cases diabetes could be predicted long before it was eventually diagnosed.

As the percentage of newly diagnosed patients querying about symptoms is very low, we were interested in assessing the rate at which internet users query for worrisome symptoms prior to diagnosis of symptomatic medical conditions. We repeated the analysis for 10 medical conditions in which noticeable worrisome symptoms appear in the majority of patients and for which there are no routine screening measures. We estimated the fraction of people who newly self-identified as having a diagnosis of one of these conditions who queried about a symptom associated with that condition in the 10 months before diagnosis. On average, 14% (s.d. 8%) of patients query for symptoms associated with these 10 conditions (Table 3).

Discussion

In this study, we identified 11,050 Bing users who mentioned they have a new diagnosis of diabetes and approximately 11.5 million users who queried for symptoms that can be attributable to diabetes. In the diabetes diagnosis group, only 377 users (3.4%) made at least one symptom query in the 10 months before diagnosis. A linear regression model was very successful in distinguishing the diabetes diagnosis users who queried about symptoms from the control users who queried about symptoms (AUC 0.89), and this was true for long time periods prior to diagnosis (up to 240 days).

We assessed the predictive value of each of the symptoms associated with diabetes for a future query indicating diabetes diagnosis. The symptom that contributed most to the prediction was malaise, followed be fatigue, and mentions of urine (in any context).

In the logistic regression model, some of the symptom query terms (fatigue, hunger, somnolence, slow healing, and weight loss) were associated with a lower likelihood of future diabetes diagnosis. We can speculate that either these symptoms are less specific or that because they are commonly recognized as symptoms of diabetes, users querying on these symptoms are more likely to have already excluded diabetes as the cause.

We were also interested in assessing the timing of queries about each of the predictive symptoms. The symptoms that were queried about shortly before the diabetes diagnosis query were blurred vision, fatigue and hunger. Queries on impotence appeared earlier prior to diabetes diagnosis query (10–30 days), while queries on weight loss began at least 2 months prior to the diabetes diagnosis query.

Repeating the same search strategy for symptom queries preceding diagnosis of other medical conditions which are not screened for, are generally highly symptomatic and diagnosed as a result of symptoms (Table 3), we found that even in these symptomatic conditions, symptom search before diagnosis is not universal and occurs in only 4–27% of patients, on average 14% (Sc.D. 8%). This result shows that the previously published survey results [8] overestimate the actual use of the internet for self-diagnosis. As expected, searches for diabetes symptoms are in the low range of this spectrum, reflecting the asymptomatic phase in which the majority of patients are diagnosed. We can roughly extrapolate that the 4% symptom search should reflect a magnitude of 24% rate of symptoms in newly diagnosed diabetes patient, a result that is similar to that found in other methods [3].

Our study is reliant on the admission of search engine users in the positive class that they have diabetes, and on the lack of admission by the users of the negative class.

Table 3 Conditions and the percentage of people who asked about their associated symptoms, as provided

Condition	Symptoms	% with symptoms
Degenerative disk disease	Back pain	15.2
COPD	Chronic cough, shortness of breath, dyspnea, recurrent pneumonia, wheezing	5.1
Menopause	Hot flash, night sweat, vaginal dryness, alopecia	11.4
Heart failure	Shortness of breath, dyspnea, chronic cough, leg edema, leg swelling, rapid weight gain, fatigue	12.9
Gout	Pain, tenderness, swelling, inflammation, redness	25.3
Rheumatoid arthritis	Pain, stiffness	27.0
Ulcerative colitis	Diarrhea	9.0
Bladder cancer	Blood in urine, hematuria	10.6
Parkinson's disease	Tremor, bradykinesia	3.6
Endometrial cancer	Discharge, bleeding	22.7
Average		14
STD		8

This represents two drawbacks. First, it is impossible to validate that all people of the positive class have diabetes. Moreover, some of the control population might also be suffering from diabetes. Second, the class of people who admit to having a medical condition is known to be biased [17]. Notwithstanding, prior work suggest these data are mostly accurate.

The results of this study demonstrate the ability of search engines to add a population-wide screening tool with a potential to identify some symptomatic patients at an earlier stage than the current medical care allows. Their fraction is 3.4%, which is admittingly not large, but given the high prevalence of new diagnosis of diabetes the actual number of people that could benefit from earlier detection is quite significant.

The screening for diabetes through internet searches has additional advantages, including lower cost, access to large populations, and possibly earlier detection. Moreover, while population screening for diabetes is performed in many countries, many people do not participate in these programs for reasons of access or time. Screening using web searches could overcome these limitations.

There is no information about which type of diabetes was diagnosed, but we expect type 1 diabetes to represent a higher fraction of symptomatic patients than their percentage in the patient population. Reaching an earlier diagnosis in these patients, even if only by a few weeks or days, may be extremely important and have clinical significance as it may be able to prevent many of the cases of diabetic ketoacidosis.

Our study adds to the body of knowledge which suggests that screening for diseases, including several types of cancer [11–13], can be performed using search queries. Thus, Internet searches could possibly have a wider use as a screening tool for multiple conditions. This capability, however, is not without its challenges, both medical and technical (how and when to approach users) and ethical (in the implementation of unsolicited diagnosis, the cost of false detections, etc.). Thus, deploying this capability requires important societal and legal discussions, in addition to solving the not insignificant medical and technical challenges of such population-wide screening.

The accuracy of diabetes detection by search engines could potentially be improved by integrating additional information that is routinely revealed by users through their searches, including information pertaining to classical risk factors such as age, body mass index, socioeconomic status and location. These factors could be integrated into an even better algorithm to enhance the specificity and sensitivity of diabetes risk stratification both in users querying about symptoms and in asymptomatic users that do not. The question of whether and how to provide this information remains open. Our paper, which adds to the existing body on disease detection through search engine queries, should help spawn the discussion in a wide forum of internet users, patients, health leaders and information technology experts.

Authors' contributions IH conceived the study, interpreted the results, and wrote the paper. EYT conceived the study, collected the data, analyzed it and wrote the paper. DD and SN assisted in analyzing the results and wrote the paper.

Availability of data and material The datasets analyzed during the current study are not publicly available due to privacy and terms of use but are available from the corresponding author on reasonable request.

Compliance with ethical standards

Conflict of interest EYT is an employee of Microsoft, owner of the Bing search engine. Work described herein was performed as part of the author's salaried employment. The authors declare no conflict of interest.

Ethical approval This study was approved by the Behavioral Sciences Research Ethics Committee of the Technion, Israel Institute of Technology.

Informed consent For this type of study formal informed consent is not required.

References

1. National Diabetes Statistics Report (2017) CDC, Alanta
2. Rodbard HW, Green AJ, Fox KM, Grandy S (2009) Trends in method of diagnosis of type 2 diabetes mellitus: results from SHIELD. Int J Endocrinol 2009:796206
3. O'Connor PJ (2006) Diabetes: how are we diagnosing and initially managing it? Ann Fam Med 4(1):15–22
4. International Diabetes Federation (2017) IDF diabetes atlas, 8th edn. International Diabetes Federation, Brussels
5. Bertuzzi F et al (2018) Teleconsultation in type 1 diabetes mellitus (TELEDIABE). Acta Diabetol 55(2):185–192
6. Di Bartolo P, Nicolucci A, Cherubini V, Iafusco D, Scardapane M, Rossi MC (2017) Young patients with type 1 diabetes poorly controlled and poorly compliant with self-monitoring of blood glucose: can technology help? Results of the i-NewTrend randomized clinical trial. Acta Diabetol 54(4):393–402
7. Yaron M et al (2019) A randomized controlled trial comparing a telemedicine therapeutic intervention with routine care in adults with type 1 diabetes mellitus treated by insulin pumps. Acta Diabetol. https://doi.org/10.1007/s00592-019-01300-1
8. Fox S, Duggan M (2013) Health online. Pew Research Center, Washington
9. Yom-Tov E, Gabrilovich E (2013) Postmarket drug surveillance without trial costs: discovery of adverse drug reactions through large-scale analysis of web search queries. J Med Internet Res 15(6):e124
10. Yom-Tov E, Borsa D, Hayward AC, McKendry RA, Cox IJ (2015) Automatic identification of web-based risk markers for health events. J Med Internet Res 17(1):e29
11. Soldaini L, Yom-Tov E (2017) Inferring individual attributes from search engine queries and auxiliary information, pp 293–301

12. White RW, Horvitz E (2017) Evaluation of the feasibility of screening patients for early signs of lung carcinoma in web search logs. JAMA Oncol 3(3):398
13. White RW, Doraiswamy PM, Horvitz E (2018) Detecting neurodegenerative disorders from web search signals. NPJ Digit Med 1:8. https://doi.org/10.1038/s41746-018-0016-6
14. Allerhand L, Youngmann B, Yom-Tov E, Arkadir D (2018) Detecting Parkinson's disease from interactions with a search engine: is expert knowledge sufficient? In: Proceedings of the 27th ACM international conference on information and knowledge management—CIKM'18, Torino, Italy, pp 1539–1542
15. Duda RO, Hart PE, Stork DG (2012) Pattern classification. Wiley, New York
16. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(1):29–36
17. Yom-Tov E (2019) Demographic differences in search engine use with implications for cohort selection. Inf Retrieval J. https://doi.org/10.1007/s10791-018-09349-2