ORIGINAL PAPER

Nagoya J. Med. Sci. **83**. 107–111, 2021
doi:10.18999/nagjms.83.1.107

# Usefulness of web search queries for early detection of diseases in infants

Shuji Yamaguchi[1], Akinari Hinoki[2], Kota Tsubouchi[1], Hizuru Amano[2,3], Akira Tajima[1] and Hiroo Uchida[2]

[1]*Yahoo Japan Corporation, Tokyo, Japan*
[2]*Department of Pediatric Surgery, Nagoya University Graduate School of Medicine, Nagoya, Japan*
[3]*Department of Pediatric Surgery, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan*

## ABSTRACT

Early detection of diseases is critical in infants. This study evaluates the usefulness of web searches in predicting diseases in order to encourage guardians to consult a doctor promptly if their children are ill. We collected six months of search queries from Yahoo! JAPAN Search between October 2016 and March 2017. Using a machine learning model, we investigated the accuracy of the search query's ability to predict the diagnosis of biliary atresia and hypertrophic pyloric stenosis. Both diseases were modeled with an accuracy of approximately 80%, and symptoms related to the disease were significant features in the model. These findings suggest the possibility of detecting diseases from web search queries performed by guardians. Through future research, we intend to propose a method that uses web search queries for early detection of these diseases by providing appropriate and timely information to support the guardians of patients.

Keywords: search engine, infancy, biliary atresia, hypertrophic pyloric stenosis

## INTRODUCTION

Advancements in artificial technology and medical measurement equipment are increasingly being used for medical data analysis. For early detection of diseases, however, currently available medical data are not sufficient as they do not include information on patients' behavior and symptoms prior to the hospital visit. Early detection of diseases is particularly critical in infants when a time-sensitive intervention is required, such as a condition known as biliary atresia that requires an operation within 60 days of onset.[1] Detecting early signs of diseases that are observable during daily activities would be beneficial in this regard. Guardians could then be encouraged to promptly consult a doctor after detecting signs of diseases. Thus, patients and their guardians always require appropriate and timely information. A web search is a promising information source that can be used to provide such information as it captures the daily activities and interests of a vast number of users. Several efforts to detect cancer using patients' own web search queries have been reported.[2,3] Since infants cannot perform a web search on their

Received: March 16, 2020; accepted: July 28, 2020
Corresponding Author: Hiroo Uchida, MD, PhD
Department of Pediatric Surgery, Nagoya University Graduate School of Medicine,
65 Tsurumai-cho, Showa-ku, Nagoya 466-8560, Japan
Tel: +81-52-744-2959, E-mail: hiro2013@med.nagoya-u.ac.jp

own physical condition, we must rely on the web search queries of their guardians for similar information.

The present study sought to evaluate the usefulness of using web search queries to predict diseases to encourage guardians to consult a doctor promptly if their infants are suspected of suffering from biliary atresia or hypertrophic pyloric stenosis. We analyzed the web search behaviors of guardians of infants who had needed prompt medical care and modeled their behaviors using machine learning techniques.

## MATERIALS AND METHODS

*Data Collection and Preparation*

We collected six months of search queries from Yahoo! JAPAN Search between October 2016 and March 2017. After obtaining consent from the users, we made the data available for our research purposes through appropriate procedures, including asking for and receiving appropriate ethical approval, following our corporate privacy policy,[4] and complying with national laws.

Infant diseases such as biliary atresia and hypertrophic pyloric stenosis are not familiar to laypeople because they occur infrequently. Therefore, the day guardians searched for biliary atresia or hypertrophic pyloric stenosis was defined as that on which their children were diagnosed with the disease. Next, we listed common symptoms of the disease and calculated their search counts by days-to-diagnosis for each disease. As common symptoms of biliary atresia, we chose "jaundice," "white stool," "vomit," and "constipation." For hypertrophic pyloric stenosis, we chose "jaundice," "vomit," and "constipation."[5,6]

*Data Analysis Strategy*

We excluded medical personnel, such as doctors, nurses, and medical students from the obtained user list who had searched the target disease name in the collected data, as mentioned above. A filtering process was then used on the data. We first listed the top 200 words that co-occurred in a search with the disease name. We then used a survey to identify unique words that were used only by medical personnel. This excluded 172 words.

We then developed predictive models for each target disease using search behaviors. We selected positive samples from the remaining users after the filtering process described above, and we selected the users who were assumed to have children (e.g., users who searched for "kid's cold") as negative samples. We used all the search words of the target period as features. We used L1-regularized logistic regression with down-sampled negatives to balance labels and tuned hyperparameters by a 10-fold cross-validation.

## RESULTS

*Usefulness of Collected Data*

Figure 1 shows the changes over time in the number of searches for common symptoms searched by people considered to be guardians of patients diagnosed with biliary atresia or hypertrophic pyloric stenosis, respectively. Day-0 represented the timing when the disease was searched, and the *X*-axis is displayed in reverse time order. The search queries included the symptoms we chose as common symptoms of the disease, which indicates the usefulness of the collected data for analysis. The number of searches for symptoms related to the disease was found to increase immediately before searching for the disease.

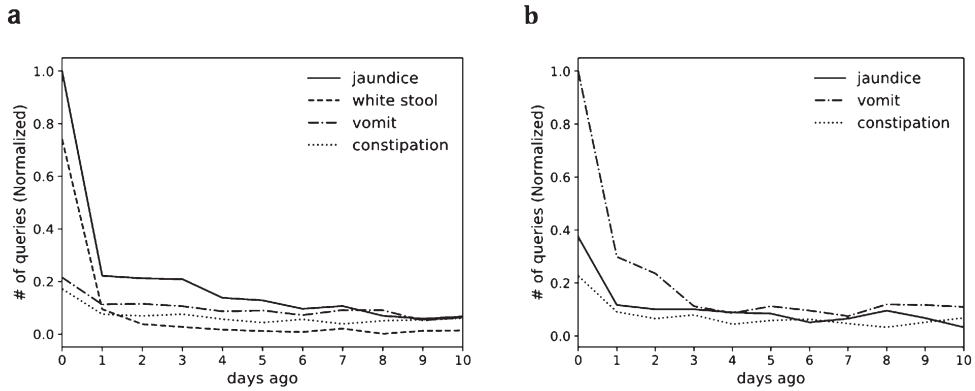


**Fig. 1** Queries about symptoms before searching for the keywords
**Fig. 1a:** Biliary atresia
**Fig. 1b:** Hypertrophic pyloric stenosis
The day in which guardians searched for keywords "biliary atresia" or "hypertrophic pyloric stenosis" was defined as being the day in which their children were diagnosed with the disease. The search counts for each common symptom by those who searched for (a) biliary atresia or (b) hypertrophic pyloric stenosis were measured by the days-to-diagnosis.

*Predicting Search Behavior Using Machine Learning*

Table 1 shows the accuracy of the models as well as the top five words in order of the most significant influence on the model's ability to predict the target disease. For biliary atresia, the top five queries were newborn/新生児, color of stool/うんちの色, jaundice/黄疸, stool/うんち, and biliary tract/胆道. For hypertrophic pyloric stenosis, the top five queries were medical/医療, groan/うなる, pyloric stenosis/幽門狭窄, how to memorize/覚え方, and Twitter/ツイッター. Both diseases were modeled with an accuracy of approximately 80%, and symptoms related to the disease (emboldened in Table 1) appeared as significant features in the model. We can predict whether or not people who search for the words listed in our predictive models for each disease (Top 5 words in Table 1) will also search for biliary atresia or hypertrophic pyloric stenosis with an accuracy of approximately 80%, which indicates that their infants may be diagnosed with the specific disease.

**Table 1** Accuracy and Top Five Queries of Models

| | Biliary atresia | Hypertrophic pyloric stenosis |
|---|---|---|
| Accuracy | 0.79 | 0.81 |
| Top five queries | Newborn | Medical |
| | **Color of stool** | Groan |
| | **Jaundice** | **Pyloric stenosis** |
| | **Stool** | How to memorize |
| | Biliary tract | Twitter |

## DISCUSSION

Insights about days to diagnosis from the first observation of symptoms could be beneficial for developing predictive models as well as incorporating medical knowledge.[7] Hospital medical records generally do not include sufficient data on patients' behavior and symptoms before hospital visits, making it difficult to garner this information from other sources. This study clarified that it is possible to detect diseases, such as biliary atresia and hypertrophic pyloric stenosis, from web search queries performed by guardians. The predictive accuracy of diagnosis based on the machine learning model (Table 1) was found to be approximately 80%, indicating that the keywords are related.

First, we confirmed the usefulness of our dataset. Our data included queries of common symptoms for the studied diseases. We found that the number of searches for symptoms was found to increase immediately before searching for the disease name. Then, we designed a filtering process to exclude personnel from user lists based on doctors' questionnaires. Since medical personnel also use web searches, the search query contained several medical terms. This bias could lead to a model that does not suit the early detection of the disease by the patient themselves, so our filtration method could have great value in detecting diseases without bias.

In biliary atresia, "newborn," "color of stool," and "jaundice" were all keywords in the web queries related to the infant's symptoms. After many web searches using various combinations of these words, guardians would hit "biliary atresia." When query words, such as "infant" and "jaundice," are searched, the study authors plan to create additional information regarding biliary atresia on the web. Future studies can help devise a method that utilizes web search queries for early detection of these diseases by providing appropriate and timely information that supports patient guardians.

On the other hand, web queries characteristics of hypertrophic pyloric stenosis were different from those of biliary atresia. "Milk" and "vomiting" were not included in the top five queries. These words, which are very common for many situations and diseases, are not specific for hypertrophic pyloric stenosis. To arrive at a result of hypertrophic pyloric stenosis, the specific keyword "pyloric stenosis" had to be included in the search terms. Hypertrophic pyloric stenosis often occurs in about the first month of life.[5,6] "Groan" is one of the characteristic events that occurs when an infant is around one month old; it is possible, therefore, that those searching for "infant groan" are parents of one-month-old infants. The combination of these two words suggests the disease of hypertrophic pyloric stenosis. The other two queries, "Twitter" and "how to memorize," are not intuitively related to the disease. "How to memorize" is probably a search phrase that students and nurses use when studying about illness and indicates that our filtering process was not completely effective and should be further improved in future work. In addition, since "Twitter" is machine-learned for all queries not limited to the medical words in our method, such daily queries should be included in the model. We believe that words that are not directly related to such diseases may also be useful in representing patient behavior. For example, an infant whose guardian searches for "how to burp a baby" may frequently vomit.

Our observations indicate directions for future research. As shown in the usefulness of the collected data survey (Figure 1), the number of symptoms related to the disease declined as it moved away from Day-0, but the attenuation rate varied. The reason for this is considered to be the seriousness of the symptom: if it is a serious symptom, a patient is usually promptly taken to the hospital, where they are diagnosed. Another reason may be the connection between certain symptoms and a specific disease, as in the color of stool and jaundice, for example. Given these variations, in further research, we must look into the effectiveness of using sequences of symptoms, such as vomiting after a high fever, as features of the prediction model, and collate

it with medical knowledge. Therefore, the importance of early detection by the prediction model is very high for infants. It also means that the prediction task is more difficult. Therefore, efforts to improve predictive accuracy using only objective symptoms remain important.

This study has a few limitations. First, we could not entirely exclude non-patients' guardians, such as medical personnel, from the user list. Furthermore, we could not access the users' medical records such as diagnosis names and data, so we conveniently defined the "diagnosis" to identify a definitive disease name on the web. However, our findings indicate the terms searched by patients' guardians, which could accelerate more targeted future research on diseases using search queries.

## CONCLUSIONS

We proposed a unique method using web search queries for the early detection of diseases to provide appropriate and timely information to support patients' guardians. It was confirmed that it is possible to detect diseases from web search queries performed by a person other than the patient. The words identified by the machine learning models were verified with the prediction accuracy of approximately 80%, although some were not directly related to the disease.

## CONFLICTS OF INTERST STATEMENT

S. Yamaguchi, K. Tsubouchi, and A. Tajima were employed by Yahoo! Japan Corporation at the time this article was written.

## REFERENCES

1 Shirota C, Uchida H, Ono Y, et al. Long-term outcomes after revision of Kasai portoenterostomy for biliary atresia. *J Hepatobiliary Pancreat Sci*. 2016;23(11):715–720. doi:10.1002/jhbp.395.
2 Paparrizos J, White RW, Horvitz E. Screening for pancreatic adenocarcinoma using signals from web search logs: feasibility study and results. *J Oncol Pract*. 2016;12(8):737–744. doi:10.1200/JOP.2015.010504.
3 Paul MJ, White RW, Horvitz E. Search and breast cancer: on episodic shifts of attention over life histories of an illness. *ACM Trans Web*. 2016;10(2):1–27. doi:10.1145/2893481.
4 Yahoo! Japan Corporation. Privacy Policy. Yahoo! Japan. https://about.yahoo.co.jp/docs/info/en/terms/privacypolicy. Published October 2019. Access June 16, 2020.
5 Amano H, Kawashima H, Iwanaka T. Pyloromyotomy. In: Taguchi T, Iwanaka T, Okamatsu T, eds. *Operative general surgery in neonates and infants*. Tokyo: Springer; 2016:185–191.
6 Li J, Celiz AD, Yang J, et al. Tough adhesives for diverse wet surfaces. *Science*. 2017;357(6349):378–381. doi:10.1126/science.aah6362.
7 White RW, Horvitz E. From health search to healthcare: explorations of intention and utilization via query logs and user surveys. *J Am Med Inform Assoc*. 2014;21(1):49–55. doi:10.1136/amiajnl-2012-001473.