Association between Search Behaviors and Disease Prevalence Rates at 18 U.S. Children's Hospitals

Dennis Daniel[1,2] Traci Wolbrink[1,2] Tanya Logvinenko[3] Marvin Harper[4,5] Jeffrey Burns[1,2]

[1] Division of Critical Care Medicine, Department of Anesthesiology, Perioperative and Pain Medicine, Boston Children's Hospital, Boston, Massachusetts, United States
[2] Department of Anesthesia, Harvard Medical School, Boston, Massachusetts, United States
[3] Clinical Research Center, Boston Children's Hospital, Boston, Massachusetts, United States
[4] Department of Emergency Medicine, Boston Children's Hospital, Boston, Massachusetts, United States
[5] Department of Pediatrics, Harvard Medical School, Boston, Massachusetts, United States

Address for correspondence Dennis Daniel, MD, Division of Critical Care Medicine, Boston Children's Hospital, 300 Longwood Avenue, Boston, MA 02115, United States
(e-mail: dennis.daniel@childrens.harvard.edu).

Appl Clin Inform 2017;8:1144–1152.

Abstract

Background Usage of online resources by clinicians in training and practice can provide insight into knowledge gaps and inform development of decision support tools. Although online information seeking is often driven by encountered patient problems, the relationship between disease prevalence and search rate has not been previously characterized.

Objective This article aimed to (1) identify topics frequently searched by pediatric clinicians using UpToDate (http://www.uptodate.com) and (2) explore the association between disease prevalence rate and search rate using data from the Pediatric Health Information System.

Methods We identified the most common search queries and resources most frequently accessed on UpToDate for a cohort of 18 children's hospitals during calendar year 2012. We selected 64 of the most frequently searched diseases and matched ICD-9 data from the PHIS database during the same time period. Using linear regression, we explored the relationship between clinician query rate and disease prevalence rate.

Results The hospital cohort submitted 1,228,138 search queries across 592,454 sessions. The majority of search sessions focused on a single search topic. We identified no consistent overall association between disease prevalence and search rates. Diseases where search rate was substantially higher than prevalence rate were often infectious or immune/rheumatologic conditions, involved potentially complex diagnosis or management, and carried risk of significant morbidity or mortality. None of the examined diseases showed a decrease in search rate associated with increased disease prevalence rates.

Conclusion This is one of the first medical learning needs assessments to use large-scale, multisite data to identify topics of interest to pediatric clinicians, and to examine the relationship between disease prevalence and search rate for a set of pediatric diseases. Overall, disease search rate did not appear to be associated with hospital

Keywords

- information retrieval
- search query log analysis
- educational needs assessment
- data mining

received
August 10, 2017
accepted after revision
October 14, 2017

DOI https://doi.org/10.4338/ACI-2017-08-RA-0137.
ISSN 1869-0327.

disease prevalence rates based on ICD-9 codes. However, some diseases were consistently searched at a higher rate than their prevalence rate; many of these diseases shared common features.

Background and Significance

The volume of medical knowledge is growing exponentially, and it is increasingly challenging for clinicians to acquire and maintain the knowledge needed to deliver safe, high-quality patient care.[1–3] Clinicians frequently rely upon online medical resources to assist with clinical decision making, to supplement self-directed learning, and to stay abreast of the latest evidence.[4–6] Clinicians report that online information-seeking behavior is often driven by the need to address knowledge gaps related to patient problems they encounter in practice.[7] However, roughly half of questions generated in the course of patient care are left unanswered.[8] To most effectively utilize the scarce time and resources available for clinicians to access knowledge bases and decision support tools, it is essential to utilize systematic approaches to identifying and prioritizing knowledge gaps that are attuned to the patient care context. Traditional methods of needs assessment include surveys, examinations, practice audits, and collection of expert and organizational opinions.[9–12] These approaches are often slow, resource-intensive, difficult to scale, prone to sampling, observer and reporting biases, and separated from natural contexts of medical practice and learning.

When clinicians access medical Web sites, their browsing and searching activities are frequently directed toward acquiring information on specific topics of interest.[13,14] These activities are automatically logged by many Web sites. Businesses and other enterprises are now regularly analyzing such user activity data to better understand the needs and preferences of their clientele.[15,16] The literature lacks published examples of using similar approaches to inform the development of medical education and decision support resources.

Objective

Accordingly, we undertook a descriptive study to identify topics of interest for pediatric practitioners from a cohort of 18 children's hospitals in the United States by examining browsing behavior on UpToDate (http://www.uptodate.com), a peer-reviewed medical knowledge and clinical decision support Web site. We chose to explore browsing behavior in UpToDate because it is widely used by physicians and residents in hospitals throughout the United States.[17] We focused on data from a cohort of children's hospitals because as pediatric educators, we wanted the results to be relevant to our own clinical and learning community, but a similar general approach could be used to identify general medical, surgical, and nursing learning needs as well.

We then selected a set of frequently searched diseases and examined whether clinician search activity related to a given disease was associated with the prevalence rate of that disease in a given hospital, using ICD-9 discharge data from the Pediatric Health Information System (PHIS). Conceptually, diseases that are searched substantially more than they are encountered may be ones for which clinical knowledge gaps are more prominent, whereas diseases for which decreased search rates are associated with higher encounter rates may be ones where clinical knowledge gaps are less significant. This may be related to characteristics of diseases themselves, effects of exposure on the development of both formal knowledge and pattern recognition, and other factors. Measuring the relationship between disease prevalence and search rate can help prioritize development of educational interventions and decision support tools in a clinically oriented and data-driven fashion.

Methods

The study protocol was reviewed and approved for exemption by the Institutional Review Board at Boston Children's Hospital.

Generation of Counts of UpToDate Search Queries, UpToDate Resources Accessed, and PHIS ICD-9 Diagnosis Codes

We identified 18 U.S. freestanding academic children's hospitals that have a unique UpToDate institutional account, and that report International Classification of Diseases, Ninth Edition (ICD-9) discharge data to the PHIS, a comparative pediatric database that includes clinical and resource utilization data for inpatient, ambulatory surgery, emergency department, and observation unit patient encounters. For each hospital in the cohort, we created two datasets. The first dataset contained all UpToDate search queries from that hospital, with the first resource accessed following submission of a search query, during the 2012 calendar year. The second dataset contained all ICD-9 codes submitted from the hospital to PHIS during the same time period. The PHIS database query returned only counts of ICD-9 codes, making it impossible to identify individual patients. All query and ICD-9 code data were aggregated at the hospital level, and individual hospitals were deidentified, so that individual users, individual patients, geographic locations, or specific hospitals cannot be identified from these data. We generated the counts for all unique UpToDate search queries, for all unique UpToDate resources (based on resource title), and for all ICD-9 codes, and ranked the respective lists in descending order.

Disease Topic Selection and Generation of Query Rates and Prevalence Rates

To generate a set of disease topics for which query and disease prevalence rates could be compared, we started with the 100 most frequently submitted search queries, and eliminated medications, nonspecific conditions, diseases or conditions that did not have corresponding ICD-9 codes, and diseases for which any hospital submitted zero associated diagnosis codes during the entire 2012 calendar year. This led to a set of 64 diseases and conditions.

We focused on the top 100 most frequent search queries for three reasons:

1. We expected that all of the diseases isolated from this set would be of interest to a wide range of clinicians, as they were all frequently searched in this dataset.
2. We anticipated that the final list of disease topics selected would be large, but not overwhelmingly so.
3. We expected that the final list of disease topics would include both common and rarely diagnosed diseases to test our hypothesis.

We excluded medications because their usage is not captured by ICD-9 codes. "Non-specific condition" refers to findings or disorders such as fever or diarrhea that are present across a large variety of disease states, which confounds the counting of related search queries and ICD-9 codes.

Multiple distinct search queries can be used to access information on the same disease topic. For example, the search queries "kawasaki," "kawasaki disease," and "kawasaki disease children" all correspond to the topic of Kawasaki disease. To accurately estimate the magnitude of clinician interest in a disease topic, we tagged all search queries in the dataset related to a given disease topic using a regular expression-driven search process, followed by manual review. The counts of all of these related search queries were then summed to generate the total query count for the disease topic.

For each disease topic, we then mapped all applicable ICD-9 codes by searching for the given topic and synonyms within the text descriptions contained in the 2012 ICD-9 Tabular List of Diseases. This yielded 473 ICD-9 codes in total. For each disease, we summed all encounters where one of the mapped ICD-9 codes was present, resulting in a total diagnosis code count for each disease. The mapping of diseases to ICD-9 codes is available in ►**Supplementary Material**, available in the online version.

Because the disease topic query and ICD-9 code counts varied among the included hospitals, we used query and ICD-9 code rates to better compare the hospitals to each other. The topic query rate for each hospital was generated by taking the topic query count for that hospital, and dividing by the total number of search queries from the given hospital. We generated topic disease prevalence rates by taking the topic ICD-9 code count and dividing by total number of ICD-9 codes from the given hospital. Both disease topic query and prevalence rates were calculated by hospital and calendar month, such that each disease topic had 216 (18 hospitals × 12 months) pairs of query and prevalence rates.

Linear Regression Analysis of Query Rates and Disease Prevalence Rates

We utilized linear regression to examine the association between disease prevalence rate and query rate, paying particular attention to the slope of the regression line.

We hypothesized:

- Disease topics with positive regression slopes greater than 1 were ones where clinician search behavior was motivated by factors beyond disease exposure.
- Disease topics with slopes close to 1 were ones where clinician search behavior may be predominantly associated with disease exposure.
- Disease topics with slopes close to 0 were ones where clinician interest and clinical exposure were not associated with each other.
- Disease topics with slopes less than 0 were ones where greater exposure was associated with decreased clinician information-seeking need.

First, to explore how topic query rate was overall associated with disease prevalence rate, we calculated a linear regression for the entire set of 64 disease topics. For each disease, we used the median disease prevalence rate as the independent variable, and the median query rate as the dependent variable. We then performed linear regression for each individual disease, using the 216 pairs of query and disease prevalence rates for each disease topic. *p*-Values were generated with the null hypothesis that the slope would equal 0.

Data were analyzed using Stata/SE 13.1 for Mac (StataCorp LP, 2014).

Results

Demographics of Included Hospitals and Search Queries

►**Table 1** presents demographic data for the hospital cohort. Data on bed capacity was obtained from the 2012 AHA Hospital Database. All included hospitals were freestanding

Table 1 Characteristics of the hospital cohort for calendar year 2012

Characteristics	Number of hospitals ($N = 18$)
Bed capacity	
300 or less	8
301–400	5
401 or greater	5
Number of PHIS encounters	
50,000–100,000	8
100,001–150,000	5
150,001 or greater	5
Number of UpToDate queries	
50,000 or less	9
50,001–100,000	6
100,001 or greater	3

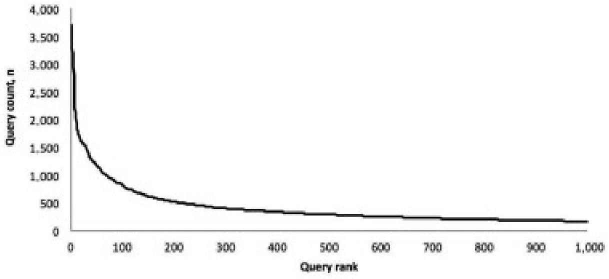

Fig. 1 Search query rank versus query count for the top 1,000 queries submitted by the hospital cohort in 2012.

children's hospitals affiliated with an academic institution. Fifteen of the 18 hospitals had 200 or more patient beds. All hospitals had resident trainees (pediatric residents and other medical trainees), and all had at least 50,000 encounters reported to PHIS.

Characteristics of Search Query and Resource Access Behavior

A total of 1,228,138 search queries across 592,454 sessions were submitted from the cohort of 18 hospitals in calendar year 2012. An UpToDate resource was accessed following 86% of all search queries (1,059,243/1,228,138). Of these, 84% (893,913/1,059,243) of queries led to a user clicking a resource providing information related to a disease or condition, and 13% (142,482/1,059,243) of queries led to clicking on information related to a medication.

Sessions contained an average of two search queries, with 54% of all sessions (322,364/592,454) containing only one search query and 95% of all sessions (563,097/592,454) containing five or fewer search queries. Forty-five percent of sessions with more than one search query contained at least one duplicated search query (121,266/270,090), and in general, search sessions were most often focused on one topic.

►Fig. 1 illustrates the distribution of search query count versus frequency rank for the top 1,000 search queries. Both search query frequency and resource access frequency followed exponential distributions in our dataset, whereby the most frequent search queries submitted (or resources clicked) accounted for a large proportion of the total. Similar distributions have been observed in other search log databases.[18]

►Table 2 lists the top 10 most frequent search queries overall from among the 18 hospitals, as well as the titles of the top 10 UpToDate resources accessed (based on counts of the first link clicked by users after a search). There is partial overlap in the disease topics represented in the two lists. As noted in Section "Disease Topic Selection and Generation of Query Rates and Prevalence Rates," multiple distinct search queries may be used to access information on the same disease topic. The overlap in disease topics between the two tables demonstrates the utility of examining counts of distinct search queries as a quick screen, while illustrating how sole reliance on such counts can lead to an incomplete understanding of user behavior.

Table 2 Top 10 UpToDate search queries and resources accessed by the hospital cohort in calendar year 2012

A. Top 10 search queries	
Query	**Count**
Kawasaki	3,707
Croup	3,577
Pertussis	3,276
UTI	3,198
Asthma	2,952
HSP	2,808
Pneumonia	2,340
Clindamycin	2,196
Bronchiolitis	2,126
Pancreatitis	2,017
B. Top 10 UpToDate resources accessed	
Resource title	Count
Acute management, imaging, and prognosis of UTIs in infants and children older than 1 mo	4,497
Kawasaki disease: clinical features and diagnosis	3,776
Dermatophyte (tinea) infections	3,230
Treatment and prevention of streptococcal tonsillopharyngitis	3,210
Acute otitis media in children: treatment	2,352
Approach to the management of croup	2,317
Clinical manifestations and diagnosis of	2,311
Conjunctivitis	2,234
Approach to the patient with abnormal liver function tests	2,227
Febrile seizures	1,971

Abbreviations: HSP, Henoch-Schönlein purpura; UTI, urinary tract infection.

Relationship between Disease Query Rate and Disease Prevalence Rate

►Fig. 2 displays the scatterplot of median query rate versus median ICD-9 code prevalence rate for the 64 included disease topics, along with the linear regression line. The regression line had a slope of 0.24 with 95% confidence interval (CI) 0.09 to 0.40 ($R^2 = 0.28$, $p < 0.004$). A table reporting the median ICD-9 code prevalence and query rates for the 64 disease topics is available in ►Supplementary Material, available in the online version. Because ICD-9 codes for asthma were submitted very frequently compared with the other included topics (median 24.8 codes per 1,000 for the overall hospital set), we also performed a regression analysis excluding the asthma data point. The regression line in this case had a similar slope of 0.58 with 95% CI 0.21 to 0.94 ($R^2 = 0.34$, $p < 0.003$).

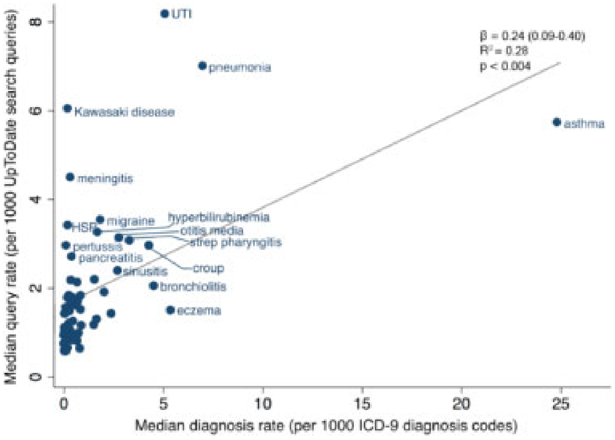

Fig. 2 Scatterplot of median UpToDate search query rate versus Pediatric Health Information System median ICD-9 code rate for 64 selected disease topics. Linear regression line displayed in gray.

►Table 3 lists the regression slopes, p-values, and R^2 values for the disease topics. None of the disease topics had a slope less than −0.20. Thirty-three of the 64 topics had regression lines with slopes between −0.20 and 0.99. Thirty-one of the topics had slopes ≥1.00, and of these, 13 topics had slopes >4.00. Many of the disease topics with slopes of 1.00 or more were infectious or rheumatologic/immune conditions (18/31, 58%).

Discussion

Previously published work in the medical literature has demonstrated how Web browsing and search data can be analyzed to understand disease trends, patterns of user interaction with Web sites, or medical topics of interest to the general population.[19–23] There are also recently published examples of using search log data from UpToDate as part of surveillance efforts related to clinician uptake of drug-safety alerts.[24] However, to our knowledge, this is one of the first published attempts at measuring pediatric knowledge needs using a large dataset of clinician activity on a popular online medical information resource. We have further attempted to contextualize these knowledge needs with data derived from clinical practice, by examining how clinician information-seeking behavior is associated with disease prevalence. Adult learning theory emphasizes the importance of ensuring that educational experiences are aligned with learners' motivations and prior experience, and that they address the needs, interests, and problems that learners currently face.[25] Our methods provide an example of how educators can leverage already-existing large datasets to better understand what medical learners want and need to know in a more precise, timely, and efficient fashion.

This approach to characterizing clinician information needs has a number of advantages. As evidenced herein, using Web browsing activity data provides a large number of data points that can be readily collected across multiple sites and geographic locations quickly and without difficulty. Online activity data are directly captured from clinicians' actions while unobserved, in natural contexts, without artificial motivators, and without need for human intervention in the data collection. Thus, the risks of observer and reporting biases, including the Hawthorne Effect (changes in behavior due to awareness of being observed), are avoided. In this effort, we analyzed all captured queries and diagnosis codes related to the selected disease topics, reducing sampling bias.

Using browsing activity data provided by UpToDate, we were readily able to generate a list of topics and resources that were frequently accessed by clinicians across a large group of U.S. hospitals. Some of the most frequently searched disease topics were commonly diagnosed, whereas others were rarely diagnosed. Overall, we did not find a consistent association between disease prevalence rate at a given hospital and corresponding search rate. However, we did identify a set of disease topics that were consistently searched at a higher rate compared with their prevalence rate, across multiple hospitals. We note that many of these diseases are difficult to diagnose without appreciating characteristic clinical findings, fulfilling defined diagnostic criteria, or performing specific laboratory or imaging evaluations. Many of these diseases also carry risk of significant morbidity or mortality, particularly if there are delays in diagnosis.

To understand why clinicians may seek certain information online at a particular time, it is important to note that clinician learning is ultimately directed to providing competent, timely, safe, and high-quality care for their patients. Much of this learning, especially in the online space, consists of informal, self-directed, and context-dependent efforts, including "just-in-time" information seeking. One theoretical framework of performance support delineates five "moments" of learning need: learning for the first time, expanding on previously acquired knowledge, attempting to recall something previously learned, troubleshooting problems, and adapting to changing circumstances.[26] Disease topics with more complex diagnostic and therapeutic considerations, potentially significant negative consequences for suboptimal identification or management, and lower levels of direct clinical experience are perhaps more likely to stimulate more moments of learning need, or moments of learning need for more individuals, motivating clinicians to seek information online. These insights may be useful for prioritizing educational and decision support efforts related to these disease topics.

We also note that the commonly encountered diseases from the top 10 most frequent UpToDate search queries—croup, urinary tract infection, asthma, pneumonia, and bronchiolitis—had slopes close to zero in our regression analysis. This suggests that for these diseases, increased disease exposure did not, in and of itself, seem to drive clinicians to more frequently seek out online information, nor did increased exposure seem to be particularly associated with decreased information need. Moreover, none of the 64 diseases we examined exhibited a substantial decrease in search rate in association with increased disease prevalence rates; the most negative slope observed was −0.20 for urinary tract infection. For individual clinicians or groups of clinicians working together, more frequent exposure to a disease could lead to repeated accession of a set of knowledge, and greater opportunity for local expertise to develop. This might reasonably be expected to lead to decreased information need. The absence

Table 3 Linear regression analysis of query rate as a function of ICD-9 code prevalence rate for the 64 selected disease topics

A. Slope (β) 1 or greater					
Disease topic	**β**	R^2	**95% CI**	***p*-Values**	**Topic category**
Rhabdomyolysis	16.37	0.304	12.14–20.60	<**0.001**	Other
Henoch-Schönlein purpura	11.13	0.322	7.55–14.71	<**0.001**	Rheum/Immune
Salmonella	9.84	0.368	6.89–12.79	<**0.001**	Infectious disease
Kawasaki disease	8.87	0.194	6.11–11.62	<**0.001**	Rheum/Immune
Eczema herpeticum	7.26	0.304	4.27–10.24	<**0.001**	Infectious disease
Pertussis	7.14	0.152	4.69–9.59	<**0.001**	Respiratory
Acute disseminated encephalomyelitis	6.14	0.101	3.23–9.05	<**0.001**	Neurology
Hemolytic-uremic syndrome	5.89	0.035	0.58–11.18	**0.030**	Rheum/Immune
Serum sickness	5.17	0.083	2.79–7.54	<**0.001**	Rheum/Immune
Mastoiditis	4.50	0.084	2.05–6.95	<**0.001**	Infectious disease
Meningitis	4.49	0.169	3.01–5.97	<**0.001**	Infectious disease
Pseudotumor cerebri	4.40	0.204	2.10–6.70	<**0.001**	Neurology
Parotitis	4.09	0.106	2.30–5.89	<**0.001**	Infectious disease
Pelvic inflammatory disease	3.65	0.149	2.28–5.03	<**0.001**	Infectious disease
Lupus	3.64	0.145	2.47–4.81	<**0.001**	Rheum/Immune
Hypercalcemia	3.23	0.032	0.81–5.65	**0.009**	Electrolyte abnormality
Pancreatitis	2.79	0.070	1.66–3.91	<**0.001**	Gastroenterology
Syndrome of inappropriate ADH secretion	2.76	0.045	1.37–4.16	<**0.001**	Endocrinology
Diabetes insipidus	2.74	0.087	1.42–4.05	<**0.001**	Endocrinology
Infantile spasm	2.68	0.148	1.89–3.48	<**0.001**	Neurology
Intussusception	2.08	0.067	1.16–3.00	<**0.001**	Gastroenterology
Erythema multiforme	1.84	0.050	0.57–3.11	**0.005**	Rheum/Immune
Orbital cellulitis	1.77	0.044	0.69–2.85	**0.001**	Infectious disease
Osteomyelitis	1.74	0.038	0.29–3.19	**0.019**	Infectious disease
Pyelonephritis	1.56	0.102	0.93–2.18	<**0.001**	Infectious disease
Mononucleosis	1.52	0.113	0.83–2.20	<**0.001**	Infectious disease
Nephrotic syndrome	1.44	0.029	0.37–2.52	**0.009**	Other
Hypernatremia	1.40	0.044	0.47–2.33	**0.004**	Electrolyte abnormality
Anaphylaxis	1.36	0.071	0.75–1.96	<**0.001**	Rheum/Immune
Respiratory syncytial virus	1.13	0.388	0.73–1.52	<**0.001**	Infectious disease
Hypocalcemia	1.00	0.015	−0.15 to 2.14	0.087	Electrolyte abnormality
B. Slope (β) less than or equal to 1					
Apparent life-threatening event	0.89	0.046	0.27–1.49	**0.005**	Other
Hyperkalemia	0.82	0.024	0.16–1.48	**0.015**	Electrolyte abnormality
Hyponatremia	0.73	0.057	0.37–1.10	<**0.001**	Electrolyte abnormality
Diabetic ketoacidosis (DKA)	0.72	0.018	0.01–1.42	**0.046**	Endocrinology
Scarlet fever	0.71	0.048	0.15–1.28	**0.014**	Infectious disease
Hyperbilirubinemia	0.70	0.128	0.36–1.04	<**0.001**	Gastroenterology
Immune thrombocytopenia purpura	0.67	0.005	−0.33 to 1.67	0.189	Hematology/Oncology
Tinea capitis	0.66	0.139	0.31–1.02	<**0.001**	Infectious disease
Pyloric stenosis	0.49	0.011	−0.13 to 1.10	0.122	Gastroenterology
Concussion	0.46	0.086	0.24–0.67	<**0.001**	Other

(*Continued*)

Table 3 (*Continued*)

B. Slope (β) less than or equal to 1					
Otitis externa	0.41	0.283	0.23–0.58	<**0.001**	Infectious disease
Migraine	0.38	0.028	0.10–0.66	**0.007**	Neurology
Croup	0.32	0.196	0.21–0.44	<**0.001**	Respiratory
Polycystic ovarian syndrome	0.32	0.002	−0.59 to 1.23	0.489	Endocrinology
Balanitis	0.28	0.011	−0.05 to 0.61	0.101	Other
Scabies	0.25	0.026	0.08–0.42	**0.004**	Infectious disease
Pneumonia	0.25	0.084	0.12–0.38	<**0.001**	Respiratory
Otitis media	0.22	0.032	0.07–0.36	**0.005**	Infectious disease
Febrile seizure	0.18	0.007	−0.12 to 0.48	0.234	Neurology
Hypokalemia	0.17	0.005	−0.08 to 0.42	0.173	Electrolyte abnormality
Sinusitis	0.16	0.065	0.05–0.26	**0.003**	Respiratory
Bronchiolitis	0.13	0.209	0.09–0.17	<**0.001**	Respiratory
Impetigo	0.10	0.006	−0.07 to 0.27	0.248	Infectious disease
Eczema	0.09	0.027	0.02–0.17	**0.014**	Rheum/Immune
Appendicitis	0.07	0.004	−0.07 to 0.21	0.316	Gastroenterology
Asthma	0.06	0.029	0.02–0.10	**0.006**	Respiratory
Failure to thrive	0.05	0.003	−0.05 to 0.15	0.296	Other
Streptococcal pharyngitis	0.04	0.004	−0.06 to 0.13	0.430	Infectious disease
Lymphadenitis	0.02	0.000	−0.54 to 0.58	0.952	Infectious disease
Cystic fibrosis	−0.02	0.000	−0.30 to 0.27	0.902	Respiratory
Urticaria	−0.09	0.003	−0.30 to 0.13	0.425	Rheum/Immune
Neuroblastoma	−0.19	0.002	−0.58 to 0.20	0.344	Hematology/Oncology
Urinary tract infection	−0.20	0.019	−0.39 to −0.01	**0.036**	Infectious disease

Abbreviation: ADH, antidiuretic hormone.
Note: *p*-Values <0.05 are displayed in bold.

of this finding in our data may imply a certain degree of informational need is maintained independent of disease exposure. Given that our search and disease prevalence data are aggregated at the hospital level, potentially contributory factors include the need for clinicians to generate and maintain a basic fund of knowledge; the presence of clinicians of differing professional background, experience, and training level; turnover of trainees and other staff; and variation in information retention over time.

Limitations

This study has several limitations. First, the search queries included in this analysis can only be identified as originating from a given hospital's institutional account. In other words, it is not possible to determine the professional backgrounds or training levels of the individuals submitting queries. However, a 2013 multisite survey conducted across 118 general hospitals found that 77% of residents, 53% of physicians, and 18% of nurses cited UpToDate as one of their preferred resources to search for information related to patient care; in fact, it was the most highly cited resource among residents.[27] This suggests that UpToDate is most frequently used by physicians, particularly physicians-in-training. Our UpToDate database also did not include any information on clinician location, or what devices they used, when they submitted searches (e.g., from the workplace vs. at home, or from mobile devices vs. computers). Such data would contribute to better understanding of how and when clinicians seek information in relation to patient encounters.

Second, our disease prevalence rates are generated using ICD-9 discharge data. These codes are assigned after an encounter has ended, which, in conjunction with our use of fully deidentified and aggregated count data, prevents us from attempting to examine searches contemporaneously with particular patient encounters. Additionally, physicians and institutions vary in whether they will apply an available ICD-9 code when a given disease or condition is present, which can lead to underestimating disease prevalence rate. We have attempted to mitigate this effect by excluding from our analysis those diseases or findings which are highly prevalent and/or nonspecific and thus likely to be underestimated by ICD-9 codes (e.g., fever, diarrhea), and by counting all available ICD-9 codes, as opposed to focusing only on primary diagnosis codes.

Finally, although search query analysis can measure what topics are of interest to clinicians, it cannot answer exactly *why* clinicians are motivated to submit particular queries in the first place. Along these lines, the influence of factors such as presence of local expertise or variations in local educational practice at each hospital is not captured by these data. Focused application of qualitative methods such as direct observation and interviewing of groups of users can provide additional insight into the motivations of searchers and the interaction of environmental and exposure factors, but such investigations require significant time and resources, and require careful design to ensure both provider and patient privacy.

Conclusion

We have illustrated the ability to generate novel insights into the needs of medical learners through analysis of large-scale data generated by pediatric practitioners. With the continued growth of Web-based medical educational efforts, and the growing reliance of clinicians on the Internet for all manner of information needs, including just-in-time learning and continuing professional development,[28] the volume of such data will only increase, and the types of data collected will become more diverse. It is increasingly important for modern educators to understand the online information-seeking behaviors as well as the expressed information needs of clinicians. The more proficient we become at using available learning data, the better we can customize educational experiences, and potentially even predict future needs.

Many in the medical profession are concerned that information technology has fallen short in realizing its promised benefits while generating a host of unintended consequences.[29] In the age of increased exposure to data analytics and data-gathering efforts everywhere on the Internet, it is essential to address issues of data responsibility head-on. These include establishing appropriate methods of data collection; defining clear processes regarding data privacy, storage, access, and ownership; and guaranteeing that intended uses of the data are appropriate. The capture and use of individual digital learning footprints, and the application of such data for performance assessment or other high-stakes evaluations, should be undertaken with care to mitigate potential risks to professional reputations, whether individual or institutional. There is an emerging literature regarding issues of data ethics in learning analytics;[30,31] however, the medical learning space, with its potential to integrate both clinician- and patient-related data, will undoubtedly raise unique and unanticipated challenges.

The literature suggests that physicians have a limited ability to accurately self-assess learning needs.[32] There is significant potential for machine-assisted insights to improve how we target and prioritize educational efforts throughout the continuum of medical learning, from medical school and residency, through continuing medical education. However, the volume of available learning data is vast, and the techniques for making that data useful are still emerging. Further work and research are necessary to learn how to most effectively leverage such data.

Clinical Relevance Statement

Clinicians regularly rely on Internet-based resources to find information they need to support their clinical practice, and data related to their browsing and searching activities are being continuously captured. These rich data are currently underutilized in guiding medical educational and decision support efforts, and in promoting greater understanding of clinician information-seeking behaviors. By analyzing large-scale search log and ICD-9 code data, we explore the search queries of pediatric clinicians, relate searches with disease prevalence data, and demonstrate the potential for quickly gaining actionable insights into topics of interest to pediatric clinicians.

Multiple Choice Question

When considering the relationship between how frequently a disease is searched in an online knowledge base such as UpToDate and how often that disease is diagnosed at a given hospital, which of the following statements is most accurate?

A. The more prevalent a disease is, the less frequently it is searched in UpToDate
B. Many diseases that are searched in UpToDate substantially more than they are encountered are infectious, rheumatologic, and immune diseases
C. All of the most frequently searched diseases searched in UpToDate are commonly encountered
D. There is a direct linear relationship between UpToDate disease search and disease prevalence

Correct Answer: The correct answer is B. In a linear regression analysis of the relationship between UpToDate disease search and disease prevalence (represented by ICD-9 code rate) at 18 U.S. children's hospitals, no consistent overall relationship was identified between how frequently diseases were searched and their prevalence at a given hospital; however, a set of diseases was identified that were searched at a substantially higher rate compared with their prevalence rate. Many of these diseases were infectious, rheumatologic, or immune in nature, and many were difficult to diagnose without appreciating characteristic clinical findings, fulfilling defined diagnostic criteria, or performing specific laboratory or imaging evaluations. Many of these diseases also carry risk of significant morbidity or mortality, particularly if there are delays in diagnosis.

Disease topics with more complex diagnostic and therapeutic considerations, potentially significant negative consequences for suboptimal identification or management, and lower levels of direct clinical experience are perhaps more likely to stimulate more moments of learning need, or moments of learning need for more individuals. These features may, in turn, motivate clinicians to seek information online. These insights may be useful for prioritizing educational and decision support efforts.

Authors' Contributions

D.D. conceptualized and designed the study, performed database generation, drafted the initial manuscript and revised the manuscript, and approved the final manuscript as submitted. T. W. reviewed and revised the manuscript and approved the final manuscript as submitted. T.L. performed the initial statistical analysis, reviewed and revised the manuscript, and approved the final manuscript as submitted. M.H. assisted with conceptualization of the study and database generation, reviewed and revised the manuscript, and approved the final manuscript as submitted. J.B. refined study design, reviewed and revised the manuscript, and approved the final manuscript as submitted. All authors approved the final manuscript as submitted and agreed to be accountable for all aspects of the work.

Protection of Human and Animal Subjects

This study was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects. The study and its component analyses were approved for exemption by the Institutional Review Board of Boston Children's Hospital.

Conflict of Interest

None.

Acknowledgment

The authors gratefully acknowledge Amy Warner and UpToDate for their support in this project by providing search log data.

References

1 Densen P. Challenges and opportunities facing medical education. Trans Am Clin Climatol Assoc 2011;122:48–58

2 Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? PLoS Med 2010;7(09):e1000326

3 Smith R. Strategies for coping with information overload. BMJ 2010;341:c7126

4 Cook DA, Levinson AJ, Garside S, Dupras DM, Erwin PJ, Montori VM. Internet-based learning in the health professions: a meta-analysis. JAMA 2008;300(10):1181–1196

5 Younger P. Internet-based information-seeking behaviour amongst doctors and nurses: a short review of the literature. Health Info Libr J 2010;27(01):2–10

6 McKnight M. The information seeking of on-duty critical care nurses: evidence from participant observation and in-context interviews. J Med Libr Assoc 2006;94(02):145–151

7 Bennett NL, Casebeer LL, Zheng S, Kristofco R. Information-seeking behaviors and reflective practice. J Contin Educ Health Prof 2006;26(02):120–127

8 Del Fiol G, Workman TE, Gorman PN. Clinical questions raised by clinicians at the point of care: a systematic review. JAMA Intern Med 2014;174(05):710–718

9 Kern DE, Thomas PA, Hughes MT. Curriculum Development for Medical Education: A Six-Step Approach. 2nd ed. Baltimore, MD: JHU Press; 2009

10 Grant J. Learning needs assessment: assessing the need. BMJ 2002;324(7330):156–159

11 Norman GR, Shannon SI, Marrin ML. The need for needs assessment in continuing medical education. BMJ 2004;328(7446): 999–1001

12 Ratnapalan S, Hilliard RI. Needs assessment in postgraduate medical education: a review. Med Educ Online 2002;7(01):4542

13 Davies K, Harrison J. The information-seeking behaviour of doctors: a review of the evidence. Health Info Libr J 2007;24(02):78–94

14 Santillana M, Nsoesie EO, Mekaru SR, Scales D, Brownstein JS. Using clinicians' search query data to monitor influenza epidemics. Clin Infect Dis 2014;59(10):1446–1450

15 McAfee A, Brynjolfsson E. Big data: the management revolution. Harv Bus Rev 2012;90(10):60–66, 68, 128

16 Henke N, Bughin J, Chui M, et al. The age of analytics: competing in a data-driven world. McKinsey Global Institute. Available at: http://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world. December 2016. Accessed December 29, 2016

17 UpToDate. About Us. Available at: http://www.uptodate.com/home/about-us. Accessed December 29, 2016

18 Silvestri F. Mining query logs: Turning search usage data into knowledge. Found Trends Inform Retriev 2010;4(1–2):12–17

19 Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009;457(7232):1012–1014

20 Carneiro HA, Mylonakis E. Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clin Infect Dis 2009; 49(10):1557–1564

21 Yang L, Mei Q, Zheng K, Hanauer DA. Query log analysis of an electronic health record search engine. AMIA Annu Symp Proc 2011;2011:915–924

22 Blechner M, Kish J, Chadaga V, Dighe AS. Analysis of search in an online clinical laboratory manual. Am J Clin Pathol 2006;126(02): 208–214

23 Spink A, Yang Y, Jansen J, et al. A study of medical and health queries to web search engines. Health Info Libr J 2004;21(01): 44–51

24 Callahan A, Pernek I, Stiglic G, Leskovec J, Strasberg HR, Shah NH. Analyzing information seeking and drug-safety alert response by health care professionals as new methods for surveillance. J Med Internet Res 2015;17(08):e204

25 Merriam SB. Andragogy and self-directed learning: pillars of adult learning theory. New Dir Adult Contin Educ 2001;2001(89): 3–14

26 Gottfredson C, Mosher B. Are you meeting all five moments of learning need. Learning Solutions Magazine 2012;18

27 Marshall JG, Sollenberger J, Easterby-Gannett S, et al. The value of library and information services in patient care: results of a multisite study. J Med Libr Assoc 2013;101(01):38–46

28 Bennett NL, Casebeer LL, Kristofco RE, Strasser SM. Physicians' Internet information-seeking behaviors. J Contin Educ Health Prof 2004;24(01):31–38

29 Wachter R. The Digital Doctor: Hope, Hype, and Harm at the Dawn of Medicine's Computer Age. New York, NY: McGraw-Hill Education; 2015

30 Steiner CM, Kickmeier-Rust MD, Albert D. LEA in private: a privacy and data protection framework for a learning analytics toolbox. J Learn Anal 2016;3(01):66–90

31 Boyd D, Crawford K. Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. Inf Commun Soc 2012;15(05):662–679

32 Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. JAMA 2006;296(09):1094–1102