INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS 83 (2014) 131–139

journal homepage: www.ijmijournal.com

The accuracy of Internet search engines to predict diagnoses from symptoms can be assessed with a validated scoring system☆

*Bennett S. Shenker**

Rutgers Robert Wood Johnson Medical School and Rutgers Robert Wood Johnson Family Medicine Residency at CentraState, United States

ARTICLE INFO

Article history:
Received 16 November 2012
Received in revised form
26 September 2013
Accepted 9 November 2013

Keywords:
Internet
Diagnosis
Search engine

ABSTRACT

Purpose: To validate a scoring system that evaluates the ability of Internet search engines to correctly predict diagnoses when symptoms are used as search terms.
Methods: We developed a five point scoring system to evaluate the diagnostic accuracy of Internet search engines. We identified twenty diagnoses common to a primary care setting to validate the scoring system. One investigator entered the symptoms for each diagnosis into three Internet search engines (Google, Bing, and Ask) and saved the first five webpages from each search. Other investigators reviewed the webpages and assigned a diagnostic accuracy score. They rescored a random sample of webpages two weeks later. To validate the five point scoring system, we calculated convergent validity and test–retest reliability using Kendall's W and Spearman's rho, respectively. We used the Kruskal–Wallis test to look for differences in accuracy scores for the three Internet search engines.
Results: A total of 600 webpages were reviewed. Kendall's W for the raters was 0.71 ($p < 0.0001$). Spearman's rho for test–retest reliability was 0.72 ($p < 0.0001$). There was no difference in scores based on Internet search engine. We found a significant difference in scores based on the webpage's order on the Internet search engine webpage ($p = 0.007$). Pairwise comparisons revealed higher scores in the first webpages vs. the fourth (corr $p = 0.009$) and fifth (corr $p = 0.017$). However, this significance was lost when creating composite scores.
Conclusions: The five point scoring system to assess diagnostic accuracy of Internet search engines is a valid and reliable instrument. The scoring system may be used in future Internet research.

1. Introduction

Few studies have investigated whether or not Internet search engines can reliably predict a diagnosis. Internet search engines have several characteristics that could be useful in making a diagnosis, however. They are able to retrieve large numbers of webpages quickly and rank the webpages based on relevance. New Internet content is added continually, so Internet search engines have access to current information. Since they retrieve webpages from many sources, Internet search engines may avoid the potential bias of a small number of individuals or groups. Alternatively, the Internet is largely unregulated and may contain both reliable and false or misleading content.

Internet search engines were not designed specifically to provide computer assisted differential diagnosis or take the place of a qualified healthcare provider. The search algorithms

☆ Supported by internal program funds. No outside sources to report.
* *Corresponding author at*: 1001 W. Main Street, Suite B, Freehold, NJ 07728, United States. Tel.: +1 732 294 2540; fax: +1 732 294 9328.
E-mail addresses: bshenker@centrastate.com, bennett.shenker@gmail.com
1386-5056/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ijmedinf.2013.11.002

used to retrieve content are proprietary and not freely available for study. Therefore, it is difficult to predict if Internet search engines are suitable to help make a diagnosis without further study.

Patients commonly review treatment information on the Internet and also search based on symptoms [1–4]. For challenging diagnoses, Internet search engines can be helpful to physicians and even permit lay persons to predict diagnoses with limitations [5,6]. However, no studies have critically evaluated the diagnostic abilities of Internet search engines independent of the Internet user's knowledge and experience.

No validated instruments have been reported to assess the diagnostic accuracy of Internet search engines. Our primary objective in this study was to validate a scoring system that can be used to define and assess the diagnostic accuracy of Internet search engines when symptoms are used as search terms. Our secondary objective was to determine if there was a difference in diagnostic accuracy based on several variables such as brand of Internet search engine and terminology of symptoms (lay vs. medical language).

2. Methods

2.1. Diagnoses and symptom search terms

The investigators created a list of diagnoses commonly encountered in a primary care setting (see Appendix) that usually present with symptoms and are frequently seen more than once a day in our experience. We prepared a complete symptom list (individual words and short phrases) for each diagnosis using medical textbooks (see Appendix). The textbooks mixed medical terminology and lay language, so we used medical dictionaries and clinical expertise to create two equivalent symptom lists in lay language and medical terminology. Whenever lay and medical terms were identical (or the medical term was archaic), we used the same word or phrase in each list.

2.2. Internet search engines

The investigators chose the Internet search engines Google (www.google.com), Bing (www.bing.com), and Ask (www.ask.com), based on their popular use.

2.3. Preparation of Internet search engine output for evaluation

A flow diagram describing process to obtain webpages from the Internet search engines is found in Fig. 1. The symptoms for the 20 diagnoses in both lay language and medical terminology resulted in 40 symptom strings. The primary investigator entered each symptom string without quotations into the three Internet search engines. The primary investigator opened the links to the first five webpages retrieved by each search and saved copies of the webpages onto a hard drive. The webpage from the first link was designated rank 1, the second rank 2, and so on through rank 5 (henceforth referred to as rank order).

A total of 600 webpages were saved (20 diagnoses × 2 sets of terminology × 3 Internet search engines × 5 webpages per search). The primary investigator hyperlinked to the saved webpages from a spreadsheet and randomized the hyperlinks using a random number generator.

2.4. The diagnostic accuracy of Internet search engine (DAISE) score

We developed five mutually exclusive categories to define diagnostic accuracy. We assigned a numerical score to each category (Table 1) ranging from 5 (highest degree of diagnostic accuracy) to 1 (lowest degree). We made the decision to assign the lowest score (1) to a webpage that was devoted entirely to describing a single, wrong diagnosis (i.e., diagnosis that did not match the symptoms used as search terms). We lumped together the webpages that did not suggest any diagnosis with those devoted to two or more wrong diagnoses not including the "true diagnosis" and assigned the next higher score (2). While the definitions for DAISE scores 5, 4, and 3 are relatively straightforward, the definitions of DAISE score 2 and 1 are more open to debate. We chose to consider a single, incorrect diagnosis to be a worse outcome than either no diagnosis at all or multiple incorrect diagnoses. We speculated that a single, incorrect diagnosis might be more misleading (i.e., more convincing) than multiple, incorrect, diagnoses. Similarly, we felt that no diagnosis at all, while an unhelpful outcome, would be less harmful than a single, incorrect diagnosis. The DAISE score could be redefined into a four point or a six point system as well. The former might increase validity at the expense of differentiation between wegpages, while the latter might have the opposite effect. Given all of these considerations, we chose the five point score as defined in Table 1.

The DAISE score had face validity based on the expertise of the investigators.

2.5. Validation of the DAISE score

Six raters were divided into two groups of three. Three raters were assigned to the lay language webpages, and three raters were assigned to the medical terminology webpages. Each rater evaluated the 300 webpages and determined a DAISE score for each webpage. Based on an expected Spearman's rho of 0.7, minimum acceptable Spearman's rho of 0.5, $\alpha = 0.5$, $\beta = 0.20$, and three raters per group, we estimated that we would need a minimum sample of 40 webpages from each group of 300 to test intra-rater reliability [7]. We increased our sample to 20% of each group of 300 webpages (60 webpages). The 60 webpages from each group were selected using a random number generator. Two weeks after evaluating the original 300 webpages, the six raters evaluated both the 20% sample of webpages from their own group and the other group. See Fig. 2 for the flow diagram depicting this process.

Convergent validity was assessed by determining the inter-rater reliability for both sets of three raters (300 webpages for each group) and the inter-rater reliability of all six raters from the two, combined 20% samples (120 webpages total).