Predicting New Diagnoses of HIV Infection Using Internet Search Engine Data

To the Editor—Internet search engine data have been demonstrated to be a powerful tool to predict outbreaks of infectious diseases such as influenza and food-borne illnesses [1–3]. This method has not been applied, however, to predict incidence of chronic infections such as human immunodeficiency virus (HIV). HIV represents an important public health problem, yet there is a significant lag between initial reporting of diagnosis to public health authorities and report of regional incidence statistics by the Centers for Disease Control and Prevention (CDC). Providing more up-to-date information on incidence could help raise public awareness and aid prevention efforts.

We examined the association between annual state-level incidence of HIV diagnosis and the volume of Internet searches for the term “HIV” from 2007 to 2010 using Google Trends [4], a publicly available search-volume tool from Google. State estimates of HIV incidence were obtained from the CDC, based on mandatory confidential reporting of new HIV diagnoses. State search-volume data from Google Trends were available only in relative, not absolute, terms. For instance, the state with the highest frequency of searches for the term “HIV” in a given year (relative to all other search terms) was reported as 100, with other states measured relative to it. In 2007, for example, Maryland had the highest frequency of searches for “HIV” relative to all other Internet search terms, while California had 72% of this frequency; Google Trends reported the rate of search for “HIV” as 100 in Maryland and 72 in California. To maintain comparability across years, we normalized search rates for “HIV” to be relative to Maryland’s rate in 2007. Results were the same regardless of which state–year pair was chosen for normalization.

We estimated the bivariate association between Internet searches for “HIV” and annual HIV incidence at the state level from 2007 to 2010. To assess how closely searches for “HIV” predicted HIV incidence, we also estimated the bivariate association for 2007–2008 only and used estimates from this model to predict state HIV incidence for 2009–2010 based on state Internet searches for “HIV” during this period. We compared predicted to actual rates in 2009–2010.

States varied considerably in Internet searches for “HIV”; in 2010, for example, the frequency of searches for “HIV” in Utah was only 40% of Maryland’s rate,
compared to California, which was 70% of Maryland’s rate. State Internet searches for “HIV” were highly correlated with state HIV incidence (correlation = 0.83, \( P < .001 \); Figure 1A). Predicted rates of HIV incidence in 2009–2010 were also highly correlated with actual state estimates in those years (Figure 1B).

From 2007 to 2010, state Internet search rates for the term “HIV” using the Google search engine were strongly correlated with new diagnoses of HIV reported by the CDC. Although our results suggest that Internet search data may augment public health efforts to track infectious disease incidence beyond influenza and food-borne illnesses, data limitations precluded us from analyzing the association between Internet search rates for “HIV” and incident diagnoses in smaller geographic units such as metropolitan statistical areas. Understanding small-area trends is important, as rates of HIV diagnosis may fluctuate considerably more in small areas than in larger regions. Our analysis highlights, however, that Internet search data may be useful in predicting incident diagnoses of diseases of large public health importance such as HIV.

**Notes**

**Author contributions.** A. B. J. had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. All authors contributed to the design and conduct of the study; data collection and management; analysis and interpretation of the data; and preparation, review, or approval of the manuscript.

**Potential conflicts of interest.** All authors: No reported conflicts.

All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

**Anupam B. Jena**, 1,2,3 **Pinar Karaca-Mandic,** 4 **Lesley Weaver,** 4 and **Seth A. Seabury**

1Department of Health Care Policy, Harvard Medical School, and 2Department of Medicine, Massachusetts General Hospital, Boston, and 3National Bureau of Economic Research, Cambridge, Massachusetts; 4School of Public Health, Division of Health Policy and Management, University of Minnesota, Minneapolis; and 5RAND Corporation, Santa Monica, California


Correspondence: Anupam B. Jena, MD, PhD, Department of Health Care Policy, Harvard Medical School, 180 Longwood Ave, Boston, MA 02115 (jena@hcp.med.harvard.edu).

**Clinical Infectious Diseases** 2013;56(9):1352–3

© The Author 2013. Published by Oxford University Press on behalf of the Infectious Diseases Society of America. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/cid/cit222