Internet Queries and Methicillin-Resistant *Staphylococcus aureus* Surveillance

Vanja M. Dukic, Michael Z. David, and Diane S. Lauderdale

The Internet is a common source of medical information and has created novel surveillance opportunities. We assessed the potential for Internet-based surveillance of methicillin-resistant *Staphylococcus aureus* and examined the extent to which it reflects trends in hospitalizations and news coverage. Google queries were a useful predictor of hospitalizations for methicillin-resistant *S. aureus* infections.

Staphylococcus aureus is the most common bacterial pathogen isolated from human infections (*1*). Methicillin-resistant *Staphylococcus aureus* (MRSA) isolates are strains constitutively resistant to β-lactam antimicrobial drugs. MRSA was initially largely confined to patients with health care exposures (*2*), but in the late 1990s, genetically distinct strains emerged and spread rapidly among healthy persons in the United States. These new strains, known as community-associated MRSA (CA-MRSA), differ epidemiologically and genetically from older strains (*2,3*). CA-MRSA strains have become the most common cause of skin infections in US emergency departments (*4*).

There is no systematic surveillance system in the United States for MRSA. The Centers for Disease Control and Prevention (CDC) tracks a limited group of infections defined as invasive through the Active Bacterial Core (ABC) surveillance system reported from 9 regions. These include MRSA infections at normally sterile sites. In a 2007 report, CDC used ABC surveillance to estimate that there were 94,000 cases and 18,650 deaths caused by invasive MRSA disease in the United States in 2005 (*5*). This report received extensive media coverage and increased public awareness of MRSA (*6*).

Recent efforts to overcome surveillance limitations, in particular delay and limited geographic coverage, have included Internet protocol (IP) surveillance. IP surveillance monitors Internet search terms related to a specific disease, assuming that greater disease activity correlates with more searches. The best known IP surveillance is Google Flu Trends (*7*), although other researchers have created additional models (*8,9*). Given the lack of comprehensive surveillance, we examined whether Google search data might productively supplement existing systems to track the changing epidemiology of MRSA infections. Because MRSA, unlike influenza, is unfamiliar to many persons, we hypothesized that Internet search activity might reflect curiosity inspired by news reports and information-seeking related to actual infections or symptoms.

Author affiliations: University of Colorado, Boulder, Colorado, USA (V.M. Dukic); and University of Chicago, Chicago, Illinois, USA (M.Z. David, D.S. Lauderdale)

DOI: 10.3201/eid1706.101451

The Study

We used the Google Trends database to obtain the proportion of all Google searches that contained the words "MRSA" or "staph." "Staph" was included because many news stories refer to MRSA as "antibiotic resistant staph." "Methicillin-resistant *Staphylococcus aureus*" was too infrequently searched to be useful. Google Trends reports search activity relative to the average number of similar queries in February 2004. We only included US searches determined from IP addresses.

We extracted counts of US newspaper, wire service, and radio and television stories mentioning "MRSA" or "staph" from the LexisNexis Academic database. We spot-checked stories with the word "staph" to confirm they were about MRSA. One event or medical publication could generate multiple news stories. We hypothesized that the volume of news coverage captured the relative effect of the story on search behavior.

We used quarterly hospital discharge data from the University HealthSystems Consortium Clinical Database, which includes >90% of US academic medical centers, to calculate the proportion of hospitalizations including an MRSA diagnosis. These data were a proxy for true MRSA incidence. We used the diagnostic code for MRSA from the International Classification of Disease, 9th Revision (V09.0). MRSA hospitalizations include CA-MRSA infections that led to hospitalization and infections that developed during a hospitalization. This database includes ≤99 codes per discharge, more than other national hospital discharge databases. The likelihood of recording an MRSA diagnosis increases with longer lists of codes because of the many concurrent conditions in complex hospitalizations. Some medical centers systematically used <99 diagnoses fields. We adjusted hospitalization rates for the maximum number of codes submitted by each medical center each year. Data after the 3rd quarter of 2008 were not included because of implementation of a nationwide coding change for MRSA.

We related quarterly variation in MRSA hospitalizations to quarterly variations in search queries and news stories in a linear regression model. Because of

the effect of the 2007 CDC report on MRSA awareness, we tested 2 indicator variables: 1 to capture the spike in search activity during the 4th quarter of 2007, and 1 to account for higher levels of search activity in subsequent quarters (*10*). These 2 indicators enable the model baseline to differ during the quarters before, during, and after the 4th quarter of 2007, while keeping the relationship between hospitalization rates and Internet searches and news counts the same during the 3 periods. All statistical analyses were performed in Stata version 10.0 (StataCorp LP, College Station, TX, USA).

Details of the model and statistical methods are available in the online Technical Appendix (www.cdc.gov/EID/content/17/6/1068-Techapp.pdf). Weekly news counts are shown in Figure 1. They range from 4 to 130 before the October 2007 peak of 719, related to the CDC report, the effect of which appears to linger. The prior peak of 130 in April 2005 was related to articles in the New England Journal of Medicine describing necrotizing fasciitis associated with MRSA and the emergence of CA-MRSA in 2001–2002 (*11,12*).

Quarterly variation in Google searches for "MRSA" and "staph" are shown in Figure 2. Search behavior changed markedly after the October 2007 publication. In addition to the spike, there was a subsequent change in the relative frequency of search term "MRSA" compared with "staph." Note that the news count peak in 2005 is not seen in the Google searches, and the peak in the Google searches in the 3rd quarter of 2006 is not apparent in the news counts.

Google queries were a useful predictor of MRSA hospitalizations and explained 33% of quarterly variation when used alone. Adding news counts to the model resulted in increasing the percentage of explained variation only modestly to 41%. The news counts were not a significant addition to the model ($p = 0.18$).

Our final model, which includes search queries and the 2 temporal indicator variables, but not the news counts, is shown in the Table. The correlation between model predictions and observed hospitalization rates was 0.93 ($p<0.001$). Although data after 2007 are insufficient for definitive comparison, a better prediction before than after the 4th quarter of 2007 is suggested (Figure 2).

Conclusions

We report an IP surveillance model for MRSA incidence. We hypothesized that news coverage for such an unfamiliar disease would strongly influence search activity. However, news coverage did not affect the relationship between search queries and hospitalization rates before the 2007 CDC report. The congruence of the Internet search activity and the hospital discharge data suggest that their temporal pattern represents the actual trend in MRSA: an increasing incidence during 2004–2007, with a suggestion

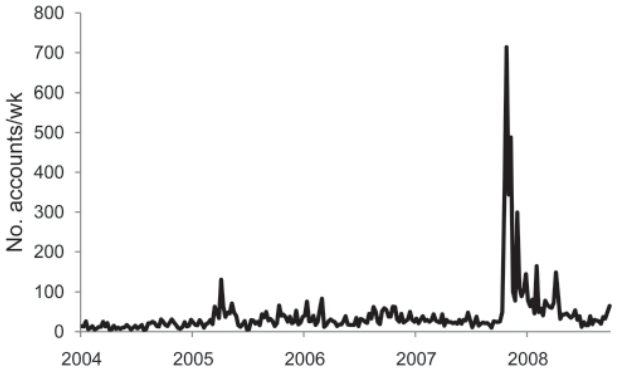

Figure 1. Weekly counts of news coverage (newspaper stories, wire service stories, and television and radio news transcripts) that mention "MRSA" (methicillin-resistant *Staphylococcus aureus*) or "staph," 2004–2008. Extracted from the LexisNexis Academic Database.

of seasonal variation, and no increase in 2008. This pattern is not the same pattern documented by the ABC surveillance data for invasive MRSA infections (*13*).

The unfamiliarity of the public with MRSA poses a challenge to using Google Trends. Searches using the phonetic misspelling "mersa" show a parallel trend to searches using "MRSA," although they are less frequent, and the correctly spelled "methicillin" is too rare to track.

Hospitalized MRSA infections include hospital-associated MRSA infections and the more serious CA-MRSA infections. Because evidence has shown that invasive hospital-associated MRSA infections decreased during the study period (*13*), the generally upward secular

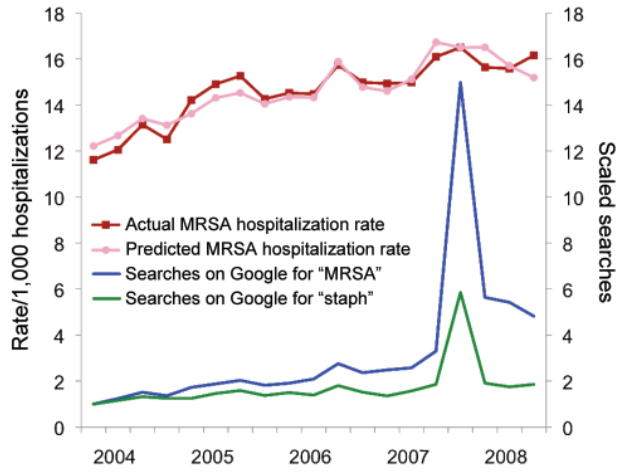

Figure 2. Actual and predicted hospitalization rates per 1,000 hospitalizations with an International Classification of Disease, 10th Revision, diagnostic code for methicillin-resistant *Staphylococcus aureus* (MRSA) and the fraction of Google search queries for "MRSA" or "Staph" (relative to the fraction of February 2004), 2004–2008.

Table. Multiple regression results for model relating UHC MRSA hospitalization rates per 1,000 hospitalizations to Google searches for "MRSA" or "staph" (normalized and scaled)*

Characteristic	Coefficient	95% CI	SE	t value	p>t
Intercept	9.03	7.56 to 10.50	0.69	13.07	<0.001
Google searches	0.25	0.18 to 0.32	0.032	7.73	<0.001
2007 4th quarter indicator	−21.45	−28.10 to −14.80	3.12	−6.87	0.001
2008 indicator	−3.06	−4.55 to −1.57	0.70	−4.37	<0.001

*UHC, University HealthSystems Consortium; MRSA, methicillin-resistant *Staphylococcus aureus*; CI, confidence interval. The overall model F(3,15) was 29.69 (p<0.0001), R^2 0.8559, and adjusted R^2 0.8270. Correlation coefficient between predicted values of this model and observed rates was 0.9251.

trend in MRSA hospitalizations is more likely to represent the trend in CA-MRSA, especially because we now know that most MRSA infections have onset in the community (*3*). The inability to distinguish community and health care infections is nonetheless a limitation of the Google and the hospitalization data. Although some hospital databases include more hospitals, they include fewer diagnostic codes. Therefore, there are no additional comprehensive data available for MRSA incidence. The lack of any true standard for MRSA incidence is why IP surveillance is potentially useful.

Acknowledgments

We thank Robert Daum for insights and numerous discussions about MRSA, Phil Schumm and Mike North for helping to quantify news coverage, and Sofia Medvedev and Samuel Hohmann for assisting with data extraction.

This study was supported by National Institute of General Medical Sciences grant U01GM087729.

Dr Dukic is an associate professor of applied mathematics at the University of Colorado, Boulder. Her research interests are Bayesian statistics, modeling of infectious diseases, sequential learning, and Internet protocol surveillance.

References

1. Lowy FD. *Staphylococcus aureus* infections. N Engl J Med. 1998;339:520–32. doi:10.1056/NEJM199808203390806
2. David MZ, Daum R. Community-associated methicillin-resistant *Staphylococcus aureus*: epidemiology and clinical consequences of an emerging epidemic. Clin Microbiol Rev. 2010;23:616–87. doi:10.1128/CMR.00081-09
3. Liu C, Graber CJ, Karr M, Diep BA, Basuino L, Schwartz BS, et al. A population-based study of the incidence and molecular epidemiology of methicillin-resistant *Staphylococcus aureus* disease in San Francisco, 2004–2005. Clin Infect Dis. 2008;46:1637–46. doi:10.1086/587893
4. Moran GJ, Krishnadasan A, Gorwitz RJ, Fosheim GE, McDougal LK, Carey RB, et al. for the EMERGEncy ID Net Study Group. Methicillin-resistant *S. aureus* infections among patients in the emergency department. N Engl J Med. 2006;355:666–74. doi:10.1056/NEJMoa055356
5. Klevens RM, Morrison MA, Nadle J, Petit S, Gershman K, Ray S, et al. Invasive methicillin-resistant *Staphylococcus aureus* infections in the United States. JAMA. 2007;298:1763–71. doi:10.1001/jama.298.15.1763
6. Hahn W, Morley C, Morrow C, Epling J. The effect of media attention on concern for and medical management of methicillin-resistant *Staphylococcus aureus*: a multimethod study. J Public Health Manag Pract. 2009;15:150–9.
7. Ginsberg J, Mohebbi M, Patel R, Brammer L, Smolinski M, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009;457:1012–4. doi:10.1038/nature07634
8. Pelat C, Turbelin C, Bar-Hen A, Flahault A, Valleron A. More diseases tracked by using Google Trends. Emerg Infect Dis. 2009;15:1327–8. doi:10.3201/eid1508.090299
9. Valdivia A, Monge-Corella S. Diseases tracked by using Google trends, Spain. Emerg Infect Dis. 2010;16:168. doi:10.3201/eid1601.091308
10. Suits D. Use of dummy variables in regression equations. J Am Stat Assoc. 1957;52:548–51. doi:10.2307/2281705
11. Miller LG, Perdreau-Remington F, Rieg G, Mehdi S, Perlroth J, Bayer AS, et al. Necrotizing fasciitis caused by community-associated methicillin-resistant *Staphylococcus aureus* in Los Angeles. N Engl J Med. 2005;352:1445–53. doi:10.1056/NEJMoa042683
12. Fridkin SK, Hageman JC, Morrison M, Sanza LT, Como-Sabetti K, Jernigan JA, et al. Methicillin-resistant *Staphylococcus aureus* disease in three communities. N Engl J Med. 2005;352:1436–44. doi:10.1056/NEJMoa043252
13. Kallen AJ, Mu Y, Bulens S, Reingold A, Petit S, Gershman K, et al. Health care–associated invasive MRSA infections, 2005–2008. JAMA. 2010;304:641–8. doi:10.1001/jama.2010.1115

Address for correspondence: Diane S. Lauderdale, Department of Health Studies, University of Chicago, 5841 South Maryland Ave, MC 2007, Chicago, IL 60637, USA; email: lauderdale@uchicago.edu

Use of trade names is for identification only and does not imply endorsement by the Public Health Service or by the US Department of Health and Human Services.

Article DOI: 10.3201/eid1706.101451

Internet Queries and Methicillin-Resistant *Staphylococcus aureus* Surveillance

Technical Appendix

Analysis

Relating temporal variation in methicillin-resistant *Staphylococcus aureus* (MRSA) diagnoses to temporal variation in search queries over time (*t)* was first conducted by using a linear regression model

$$H_t = b_0 + b_1 Q_t + b_2 N_t + e_t \quad (1)$$

where H_t denotes the quarterly proportion of hospitalization discharges from 2004 to 2008 that contained a diagnostic code for MRSA, and Q_t denotes the fraction of Google search queries that contained the words “MRSA” and/or “Staph.” Because MRSA, in contrast to influenza, is unfamiliar to many persons, we hypothesized that Internet search activity might reflect curiosity inspired by news reports and information-seeking related to actual individual infections or symptoms. To test this hypothesis, we include a second predictor, N_t, the quarterly counts of news stories extracted from the LexisNexis database.

In the above model, the coefficient b_0 denotes the average MRSA hospitalization discharge rate in the absence of any MRSA-related Internet search and news activity. Conversely, b_1 and b_2 denote the adjusted effects of the search activity and news stories, respectively, on the average quarterly MRSA hospitalization discharge rates. The error term, e_t, denotes the residual random variation in the hospitalization rates that is uncorrelated with the quarterly Internet search activity or news counts. The errors might be correlated over time, and we discuss this issue specifically toward the end of this Appendix.

Because of the effect of the 2007 Centers for Disease Control and Prevention report on awareness of MRSA among the general public, we also consider adding 2 indicator variables to the model: one to eliminate the large spike in search activity during the 4th quarter of 2007

because of this report, and one to account for the possibly increased baseline search activity level as a result of this report in the subsequent quarters. The expanded linear regression model is then

$$H_t = c_0 + c_1 Q_t + c_2 N_t + c_3 I_t + c_4 P_t + e_t \quad (2).$$

The indicator I_t is on (set to 1) only in the last quarter of 2007, and P_t is set to 1 only in post-2007 quarters. These 2 indicators enable the model baseline to differ during the quarters before, during, and after the 4th quarter of 2007. However, this model keeps the relationship between MRSA Internet searches and hospitalization rates, and between news counts and hospitalization rates, the same during the 3 periods. Thus, the coefficient c_1 describes the basic relationship between Internet activity and MRSA incidence, adjusted for the media effect stemming from news stories about MRSA and the 2007 Centers for Disease Control and Prevention report.

Finally, as with any data recorded over time, autocorrelation of the errors over time might be a potential problem. To test for the presence of autocorrelation, a Prais-Winsten transformed regression was conducted simultaneously, but little difference between the original model and the transformed model was observed. The 2 models were qualitatively similar and had nearly identical predicted values. Because our purpose focused on prediction of MRSA incidence, and not inference, we chose to present the simpler model.

Supplementary Results

The model presented in the Table in the main text of the paper shows that the effects of the 2 indicators are estimated to be negative. Had we omitted the 2 indicators, the predicted hospitalization rate in those quarters would have been high because of the high level of search activity. Thus, negative coefficients during the spike quarters eliminate the overestimate of MRSA incidence in those periods.

For completeness, we also present the Table in this technical appendix, which shows results from the model that also include news counts in addition to the predictors from the Table in the main text. This model resulted in similar statistics (overall $F[4,14]$ 20.87, $p<0.001$, adjusted R^2 0.815), increasing the correlation between the model predictions and the observed hospitalization rates negligibly from 0.9251 to 0.9254.

Technical Appendix Table. Multiple regression results for model relating UHC MRSA hospitalization rates per 1,000 persons to Google searches for “MRSA” or “staph (normalized and scaled)*

Characteristic	Coefficient	Standard error	*t* value	p>*t*	95% CI
Intercept	9.04	0.71	12.65	<0.001	(7.50 to 10.57)
Google searches	0.25	0.04	6.24	<0.001	(0.16 to 0.33)
News counts	0.0003	0.0012	0.23	0.825	(–0.002 to 0.003)
2007 4th quarter indicator	–21.66	3.37	–6.43	0.001	(–28.88 to –14.44)
2008 indicator	–3.05	0.72	–4.22	<0.001	(–4.60 to –1.50)

*UHC, University HealthSystems Consortium; MRSA, methicillin-resistant *Staphylococcus aureus*; CI, confidence interval. The overall model F(4,14) was 20.87 (p<0.0001), R^2 0.8564, and adjusted R^2 0.8154. Correlation coefficient between predicted values of this model and the observed rates was 0.9254.