Forecasting Emergency Department Visits Using Internet Data

Andreas Ekström, M Ed; Lisa Kurland, MD, PhD; Nasim Farrokhnia, MD, PhD; Maaret Castrén, MD, PhD; Martin Nordberg, MD*

*Corresponding Author. E-mail: martin.nordberg@sodersjukhuset.se.

Study objective: Using Internet data to forecast emergency department (ED) visits might enable a model that reflects behavioral trends and thereby be a valid tool for health care providers with which to allocate resources and prevent crowding. The aim of this study is to investigate whether Web site visits to a regional medical Web site, the Stockholm Health Care Guide, a proxy for the general public's concern of their health, could be used to predict the ED attendance for the coming day.

Methods: In a retrospective, observational, cross-sectional study, a model for forecasting the daily number of ED visits was derived and validated. The model was derived through regression analysis, using visits to the Stockholm Health Care Guide Web site between 6 PM and midnight and day of the week as independent variables. Web site visits were measured with Google Analytics. The number of visits to the ED within the region was retrieved from the Stockholm County Council administrative database. All types of ED visits (including adult, pediatric, and gynecologic) were included. The period of August 13, 2011, to August 12, 2012, was used as a training set for the model. The hourly variation of visits was analyzed for both Web site and the ED visits to determine the interval of hours to be used for the prediction. The model was validated with mean absolute percentage error for August 13, 2012, to October 31, 2012.

Results: The correlation between the number of Web site visits between 6 PM and midnight and ED visits the coming day was significant (r=0.77; P<.001). The best forecasting results for ED visits were achieved for the entire county, with a mean absolute percentage error of 4.8%. The result for the individual hospitals ranged between mean absolute percentage error 5.2% and 13.1%.

Conclusion: Web site visits may be used in this fashion to predict attendance to the ED. The model works both for the entire region and for individual hospitals. The possibility of using Internet data to predict ED visits is promising. [Ann Emerg Med. 2014;■:1-7.]

Please see page XX for the Editor's Capsule Summary of this article.

0196-0644/$-see front matter

http://dx.doi.org/10.1016/j.annemergmed.2014.10.008

INTRODUCTION

Background

Predicting the number of emergency department (ED) visits within any given timeframe is difficult.[1] A mismatch between ED attendance and staffing may lead to ED crowding. This is associated with decreased ED performance, as well as decreased patient satisfaction, and affects the transition from the ED to inpatient floor.[2-5] Mathematical models with which to predict the numbers of ED presentations have been developed.[1,6] Previous models commonly used linear regression or time series models,[1,6] including calendar and meteorological data, as prediction factors. The day of the week has been shown to be the strongest predictor of ED visits.[1] Several prediction models about the number of patient visits predict long-term trends and thereby are unsuitable for day-to-day adjustment of staff.[1,7]

Using Internet data for predicting the occurrence of both communicable and noncommunicable diseases has been shown to be reliable, eg, influenza outbreaks, and kidney stone occurrence.[8-14] The Swedish Institute for Communicable Disease Control has an automated surveillance system that uses Internet queries on the medical Web site the Stockholm Health Care Guide for epidemiologic surveillance.[10] The use of Internet data has also proven to be an effective tool with which to predict outcome in other fields, eg, the stock market and book and movie sales.[15-17] Internet use is extensive in Sweden, Swedish inhabitants use the Internet to the same extent as they do television,[18] and more than 90% of Swedes aged between 16 and 74 years use the Internet at least once a week.[18] Thus, Internet data may be a feasible surveillance tool for emergency medicine also.

Importance

A model with which to predict ED visits would enable better matching of personnel scheduling and ED visits and

Editor's Capsule Summary

What is already known on this topic

Several retrospectively created predictive models exist to predict future demand for emergency services.

What question this study addressed

Does use of traffic measurements from 6 PM to midnight on a regional Web site, the Stockholm Health Care Guide, correlate to emergency department (ED) utilization the following day?

What this study adds to our knowledge

The Web site activity model underestimated the actual number of ED visits but there was correlation between them.

How this is relevant to clinical practice

Targeted Internet analytic tools may help to predict specific ED utilization. Whether this model can be translated to other health care systems or regions is unknown.

thus increase ED throughput to avoid crowding, which would in turn increase patient safety.[1,7] Earlier work in the field has relied on a few independent variables. In this study, we aimed to use the predictive potential of people's behavior on the Internet by using Web site visits in an attempt to find additional factors for predicting ED inflow. Internet data could be monitored in near real time and sudden and transient changes in peoples' behavior could thereby be measured and used to predict ED visits before such changes are noticed in the ED.

Goals of This Investigation

Our hypothesis is that real-time Internet data can be used to predict ED visits. The goal of this study was to investigate whether Web site visits to an online health care guide (Stockholm Health Care Guide) could be used to predict ED attendance for the next day.

MATERIALS AND METHODS

Study Design

Models for forecasting the daily number of ED visits were derived in a retrospective, observational cross-sectional study and validated with aggregated data collected from clinical information systems and online Web site statistic tools. Data were collected for the period August 13, 2011, to October 31, 2012.

STUDY SETTING AND SELECTION OF PARTICIPANTS

The study was conducted in Stockholm County, Sweden, with a population of approximately 2 million inhabitants and a total of approximately 700,000 ED visits distributed among 7 hospitals. Stockholm Health Care Guide was the medical Web site used in the study, is operated by the Stockholm County Council, and is one of Sweden's largest Web resources for medical information (http://www.vardguiden.se).

Approximately 20% of the patients attending the ED are triaged as high-priority cases, and approximately an additional 20% seek medical attention because of wounds, fractures, or other unforeseeable injuries. Thus, approximately 60% of the cases in the ED in our setting are composed of patients with needs that are not dependent on immediate care, so they have time to seek medical information online. This is true in our setting, as well in others.[19-23]

Data Collection and Processing

The number of ED visits was collected as anonymized outcome data from the patient ledger at hospital B ED, AkuSys (version 5.0f; hospital B, Stockholm, Sweden) and from the Stockholm County Council data warehouse VAL (Stockholm regional health care data warehouse). The VAL database is a comprehensive one for all health care providers within the Stockholm County region, containing detailed health care statistics (such as date and type of visit, health care provider identification, and diagnosis at the given visit) down to the level of individual patients. All contacts with all health care providers and the corresponding diagnoses for each visit are stored in VAL, with the exception of those for a few private clinics that operate without subsidies in the Stockholm area.[24] VAL has coverage of more than 99% of hospital care in the county.[24]

The number of patient visits to the ED each hour for hospital B for the study timeframe was collected from AkuSys. The number of patient visits to the EDs each day for hospitals A, C, D, E, and F for the study timeframe was collected from the VAL database. All somatic ED visits (ie, including adult, pediatric, and gynecologic) recorded in AkuSys or the VAL databases for the given timeframe were included in the current study.

Data for visits to the Web site were collected for each hour, using the Web site statistic tool Google Analytics. All visits to the Web site were included, except those that could not be registered by Google Analytics (eg, if the Web browser was set to block cookies). The definition of a visit was the same as the one used by Google Analytics. The data were collected for the period of August 13, 2011, to October 31, 2012.

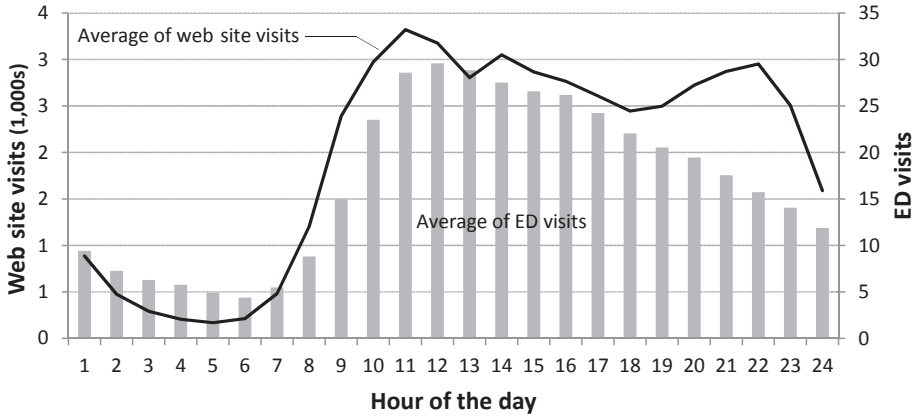

Figure 1. Mean daily variation of Web site and ED visits at hospital B. Mean number of visits per hour during a day for the training set.

Outcome Measures and Primary Data Analysis

Linear regression has been commonly used for deriving a prediction model for ED visits and has been regarded as the standard in comparing different methods.[7] Although linear regression models have been shown inferior to autoregressive integrated moving average models, for example, we chose the linear model because the current study has independent variables that have never been used in modeling. Therefore, we aimed to test our hypothesis with the most transparent method.

All data were divided into 2 sets. The training set contained data from August 13, 2011, to August 12, 2012, and the validation set contained data from August 13, 2012, to October 31, 2012. The training set was graphically analyzed with respect to variation over time and for visual patterns. The variation on a week-to-week basis was analyzed for the total number of ED visits in the county, using ANOVA. Data from hospital B, the largest hospital in the county, were used to study the hourly variation for ED visits because the VAL database contains information on the number of ED visits per day, not per hour. The analysis of hourly variation showed that both Web site and ED visits followed the same pattern, except that the Web site visits had a second peak during the evening (Figure 1), which led to the assumption that the evening visits to the Web site contained a predicting factor for next-day ED visits. Therefore, visits to the Web site during the evening (6 PM to midnight) were used to predict next-day visits to the ED. Prediction models for the number of next-day visits to the ED were derived with linear regression analysis, with the number of evening Web site visits and the day of the week as independent variables. This was conducted for each hospital, as well as for the entire county.

The linear regression model was used to predict the number of ED visits during the validation period. The predicted number of visits was compared with the actual number measured in the validation set. The mean absolute percentage error was calculated to evaluate the performance of the model and was chosen because of its common use for this purpose in previous studies and for its effectiveness in describing the accuracy of prediction models.[1,7,25] A mean absolute percentage error value below 10% was considered as good performance[1] because this would be comparable with other methods already presented. Mean absolute percentage error was calculated as the mean of all absolute differences between predicted and measured ED visits, expressed as a ratio in relation to the measured values.

SPSS (version 20; SPSS, Inc., Chicago, IL) was used. The level of significance was set to 95%, ie, $P<.05$.

All data were anonymous and aggregated, and analyses were performed on group level; thus, individuals could not be identified. No studies of patient records were performed. The data used in this study are open to the public by the Swedish principle of public access to official documents, and no ethical permit was required.

RESULTS

ED visits for Stockholm County and Web site visits to the Stockholm Health Care Guide between August 13, 2011, and August 12, 2012 (the training set), are presented in Figure 2. A seasonal variation with the 3 lowest numbers of visits around Christmas, New Year, and midsummer holidays was observed both for the ED visits in Stockholm County and the Web site visits. A weekly variation with the highest number of ED visits on Mondays and fewest visits during weekends was observed ($P<.001$). The hourly variation of ED visits to the largest emergency hospital in the region and Web site visits are presented in Figure 1. The number of ED visits had a peak at noon and then slowly decreased during the rest of the day. The hourly variation for Web site visits differed from the ED visits in that there was a second peak during the evening (Figure 1).

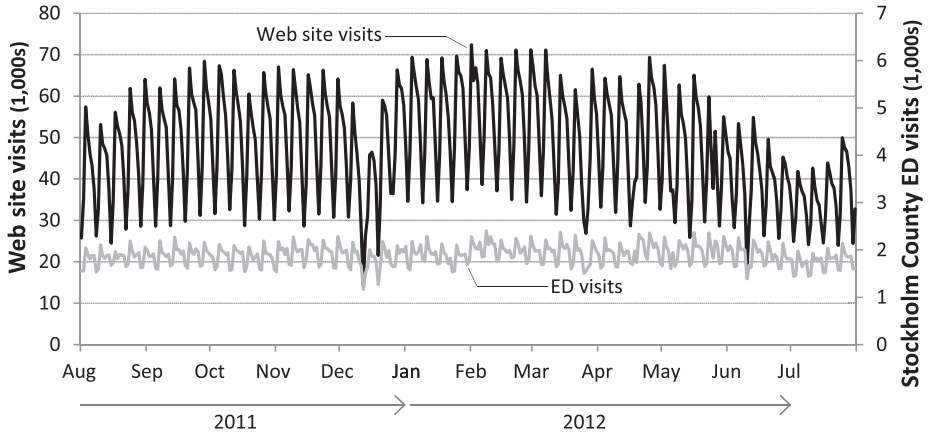

Figure 2. Number of visits to the ED in Stockholm County and Stockholm Health Care Guide Web site in the training set.

The result of the linear regression analyses with day of the week and evening Web site visits as independent variables and ED visits for each hospital individually and the county as a whole as dependent variable are presented in Table 1. The contribution of evening Web site visits as a predictor was statistically significant for all hospitals and the entire county. The correlation coefficient between ED visits in the county and evening Web site visits was 0.77 ($P<.001$) and is shown in Figure 3.

The prediction model derived was in the form

$$N = \alpha + \beta x + \sum_{i=1}^{7} \gamma_i Z_i$$

where N is the predicted number of next-day ED visits, α is a constant, β is the coefficient of the Web site evening visits (Table 1), x is number of Web site evening visits, γ is the vector of coefficients of the weekday, and z is the vector of the day of the week.

The predicted ED visits were compared with the observed visits for the same period (validation set), and mean absolute percentage error was calculated (Table 2). The best results are achieved for the entire county and the largest hospital.

Table 1. Regression analysis for ED visits.*

Dependent Variable, ED Visits, Hospital	Adjusted R^2	Coefficient for Web Site Visits, β
Stockholm County	0.779	.031
A	0.470	.004
B	0.670	.011
C	0.536	.007
D	0.725	.005
E	0.518	.004
F	0.393	.001
G	0.166	-.002

*Web site visits and day of the week were independent variables in the regression model.

Predicted and observed values for ED visits are shown in Figure 4 and Figure E1. The predicted visits follow the pattern of the measured visits but are commonly lower.

LIMITATIONS

Data were acquired August 11, 2011, to avoid a possible effect of change in the method for visitor calculation, which was performed on this date by Google Inc. The intention was to include all Web site visits to the Stockholm Health Care Guide during the study period. However, no information can be registered if the user sets the browser to block cookies. The effect of this on the current study has not been possible to measure. Nevertheless, a US study on Web privacy in 2010 showed that only 3% of Internet users block first-party cookies (used by Google Analytics).[26] We therefore assume that the number of people blocking cookies is small compared with the total number of visits to the site and therefore should not have affected the results. Another limitation was the selection of the period, ie, using Web site visits during 6 hours in the evening for next-day predictions. Perhaps a better result could have been achieved if the Web site visits for a different period had been used. The current model has not taken subgroups of ED visits into consideration, eg, different age groups or triage groups. Families with children are more likely to use the Web for health-related searches than others[18]; thus, limiting the model to predicting pediatric ED visits may prove interesting.

The models depend on the traffic to a single Web site, Stockholm Health Care Guide, and are therefore vulnerable to other Web sites competing in the same area of interest. This is a Swedish study conducted in a single region in an urban environment and in a health care system in which all citizens are included in the health care insurance and have access to the hospital ED. In addition,

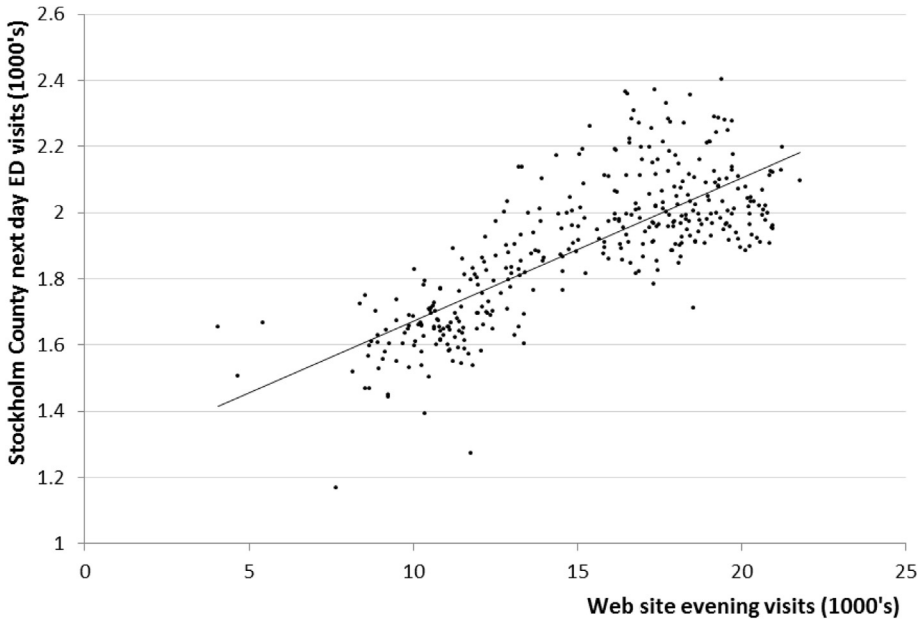

Figure 3. Correlation between evening Web site visits and next-day ED visits in Stockholm County; for the training set, r=0.77 (95% confidence interval 0.72 to 0.81).

it was conducted in a country with high Internet use in all age groups. The results may therefore be limited to similar settings. The models have not been validated for real-time use.

DISCUSSION

The current study supports that it is possible to predict ED visits by using Internet data and is, to our knowledge, the first study to use it to predict ED visits in either a region or a single hospital. We modeled predictions for ED visits for an entire region, as well as individual hospitals. The best result when predicting next-day ED visits for an individual hospital was observed in the model for the largest hospital in the region. This could imply that a high patient flow is needed for the model to be accurate, which is supported by the prediction of visits for the entire county being superior to that of any individual hospital. We believe that this study demonstrated proof of principle that it is possible to predict ED visits according to Internet data.

Table 2. Mean average percentage error for the individual hospitals and the entire county.

Hospital	ED Visits/Day*	County Correlation*	Mean Absolute Percentage Error (Max Error %)†
Stockholm County	1,895	1	4.8 (12)
A	248	0.82	6.5 (19)
B	441	0.88	5.2 (14)
C	348	0.87	9.5 (23)
D	320	0.93	7.3 (17)
E	278	0.84	6.1 (17)
F	91	0.72	8.7 (34)
G	73	0.26	13.1 (65)

*ED visits=mean number of ED visits per day during the training period. County correlation=the correlation coefficient between number of visits to a hospital ED and total number of ED visits in the county the same day during the training period.
†Minimum error for every model was 0%.

The current study shows that accurate predictions of ED visits can be made using Web site visits combined with calendar data and also when compared with studies based on regression models including only weather and calendar variables or time series models.[1,25,27,28] Reis and Mandl[27] used the time series model autoregressive integrated moving average to forecast ED daily visits in a large data set during 8 years (1992 to 2000) to trim the model and obtain a mean absolute percentage error of 9.37%. Kam et al[28] used several variables reflecting calendar and weather in a multivariate model to predict daily ED visits. Their best model had a mean absolute percentage error of 7.4%. The current model performs in a comparable fashion, with a mean absolute percentage error ranging between 5.2% and 13.1% when used to predict visits to a single hospital ED. This is also comparable with the results obtained by Jones et al[7] when they compared the ability of 6 different models to predict future visits to the ED, with mean absolute percentage errors between 8.2% and 15%, depending on model and hospital. The best mean absolute percentage error for a daily forecasting model, achieved by Wargon et al[25] by using linear regression with calendar variables,

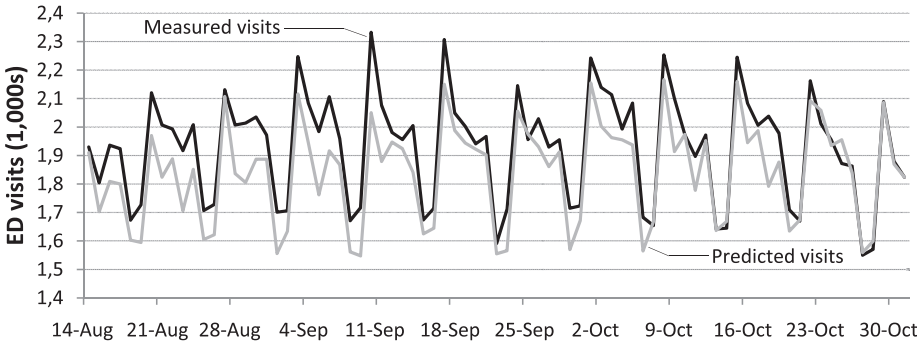

Figure 4. Predicted and measured values for all ED visits in Stockholm County. Number of ED visits per day during the validation period. ED visits declared in thousands.

showed a mean absolute percentage error of 5%. That result was calculated for the sum of visits per day for 4 hospitals in the same region. Looking at each hospital in that study, mean absolute percentage error ranged from 8.1% to 17%. The current model is as efficient as previous models, as measured by mean absolute percentage error, in predicting daily ED visits.[1,7,25,27-29] However, previous models assume a stationary process.[1] The potential benefit of the current model is that it is based on information that reflects ongoing behavioral trends and therefore makes it possible to adapt to sudden changes in behavior as a consequence or reflection of changes in the surrounding society, and thus has the potential to better reflect changes in visitor patterns to the ED.

A clinical forecasting model for short-term predictions needs static data (eg, calendar variables) or easily collected variable data. The current model used only such data, and using Internet data it is possible to continually direct and automatically collect data. This makes the models suitable for clinical implementation and surveillance.

The day of the week has previously shown to be a strong predictor for ED visits[1] and was therefore included in the current model. Still, the contribution of Web site visits was significant for all hospitals and the county as a whole and could be the reason for our models' somewhat better performance than that of previous regression models based on calendar data alone,[1] although the performance needed to successfully manage staffing is yet to be determined.

In summary, the current study provides proof of principle that Internet data can be used to predict ED visits. This is promising. For this type of information retrieval to be useful, researchers must be able to predict ED visits farther in the future than the next day. This may be possible by further investigating the correlation between Web site statistics and ED visits, and by including other variables in the model, eg, primary care visits.

The authors acknowledge Gunnar Ljunggren, MD, PhD, Stockholm County Council medical advisor, and Linda Eriksson, Web editor at Stockholm Health Care Guide, for providing us with visitor data for the ED and the Web site.

Supervising editors: Daniel A. Handel, MD, MPH; Judd E. Hollander, MD

Author affiliations: From the Department of Clinical Science and Education, Södersjukhuset, Section of Emergency Medicine, Karolinska Institutet, Stockholm, Sweden (Ekström, Kurland, Farrokhnia, Castrén, Nordberg); and the Department of Emergency Medicine, Södersjukhuset, Stockholm, Sweden (Ekström, Kurland, Farrokhnia, Nordberg).

Author contributions: AE and MN conceived the study, undertook recruitment of participating centers, and managed the data. MC obtained research funding. AE drafted the article, and all authors contributed substantially to its revision. MN takes responsibility for the paper as a whole.

Funding and support: By *Annals* policy, all authors are required to disclose any and all commercial, financial, and other relationships in any way related to the subject of this article as per ICMJE conflict of interest guidelines (see www.icmje.org). The authors have stated that no such relationships exist.

Publication dates: Received for publication April 15, 2014. Revision received September 5, 2014. Accepted for publication October 7, 2014.

Presented at the Mediterranean Emergency Medicine Congress VII, September 2013, Marseille, France.

REFERENCES

1. Wargon M, Guidet B, Hoang TD, et al. A systematic review of models for forecasting the number of emergency department visits. *Emerg Med J*. 2009;26:395-399.
2. Cowan RM, Trzeciak S. Clinical review: emergency department overcrowding and the potential impact on the critically ill. *Crit Care*. 2005;9:291-295.
3. Hwang U, Richardson L, Livote E, et al. Emergency department crowding and decreased quality of pain care. *Acad Emerg Med*. 2008;15:1248-1255.
4. Pines JM, Pollack CV, Diercks DB, et al. The association between emergency department crowding and adverse cardiovascular outcomes in patients with chest pain. *Acad Emerg Med*. 2009;16:617-625.
5. Thompson S. Efficient short-term allocation and reallocation of patients to floors of a hospital during demand surges. *Operat Res*. 2009;(57):261-273.

6. Hoot NR, Epstein SK, Allen TL, et al. Forecasting emergency department crowding: an external, multicenter evaluation. *Ann Emerg Med*. 2009;54:514-522.e19.
7. Jones SS, Thomas A, Evans RS, et al. Forecasting daily patient volumes in the emergency department. *Acad Emerg Med*. 2008;15:159-170.
8. Carneiro H. Google Trends: a Web-based tool for real-time surveillance of disease outbreaks. *Clin Infect Dis*. 2009;49:1557-1564.
9. Willard SD, Nguyen MM. Internet search trends analysis tools can provide real-time data on kidney stone disease in the United States. *Urology*. 2011;81:37-42.
10. Hulth A, Rydevik G. GET WELL: an automated surveillance system for gaining new epidemiological knowledge. *BMC Public Health*. 2011;11:252.
11. Ginsberg J, Mohebbi MH, Patel RS, et al. Detecting influenza epidemics using search engine query data. *Nature*. 2009;457:1012-1014.
12. Polgreen PM, Chen Y, Pennock DM, et al. Using Internet searches for influenza surveillance. *Clin Infect Dis*. 2008;47:1443-1448.
13. Willard SD, Nguyen MM. Internet search trends analysis tools can provide real-time data on kidney stone disease in the United States. *Urology*. 2013;81:37-42.
14. Dugas AF, Hsieh YH, Levin SR, et al. Google Flu Trends: correlation with emergency department influenza rates and crowding metrics. *Clin Infect Dis*. 2012;54:463-469.
15. Mishne G, Glance N. Predicting movie sales from blogger sentiment. AAAI 2006 Spring Symposium, March 27-29, 2006, Palo Alto, CA.
16. Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market. *J Comput Sci*. 2011;1-8.
17. Gruhl D, Guha R, Kumar R. The predictive power of online chatter. Proceedings of the KDD. New York, NY: ACM; 2005:78-87.
18. Statistics Sweden. Privatpersoners användning av datorer och internet 2011 [Peoples' computer and Internet use 2011]. 2011.
19. Lim ME, Worster A, Goeree R, et al. Simulating an emergency department: the importance of modeling the interactions between physicians and delegates in a discrete simulation. *BMC Med Inform Decis Mak*. 2013;13:59.
20. Wiler JL, Poirier RF, Farley H, et al. Emergency Severity Index triage system correlation with emergency department evaluation and management billing codes and total professional charges. *Acad Emerg Med*. 2011;18:1161-1166.
21. Christ M, Grossmann F, Winter D, et al. Modern triage in the emergency department. *Dtsch Arzebl Int*. 2010;107:892-898.
22. Ng CJ, Hsu KH, Kuan JT, et al. Comparison between Canadian Triage and Acuity Score and Taiwan Triage System in emergency departments. *J Formos Med Assoc*. 2010;109:828-837.
23. Asaro PV, Lewis LM. Effects of a triage process conversion on the triage of high-risk presentations. *Acad Emerg Med*. 2008;15:916-922.
24. Wändell P, Carlsson AC, Wettermark B, et al. Most common diseases diagnosed in primary care in Stockholm, Sweden, in 2011. *Fam Pract*. 2013:1-8.
25. Wargon M, Casalino E, Guidet B. From model to forecasting: a multicenter study in emergency departments. *Acad Emerg Med*. 2010;17:970-978.
26. Wills CE, Zeljkovic M. A personalized approach to Web privacy: awareness, attitudes and actions. *Informat Manage Comput Security*. 2011;19:53-73.
27. Reis B, Mandl K. Time series modeling for syndromic surveillance. *BMC Med Inform Decis Mak*. 2003;11:1-11.
28. Kam HJ, Sung JO, Park RW. Prediction of daily patient numbers for a regional emergency medical center using time series analysis. *Healthc Inform Res*. 2010;16:158-165.
29. Batal H, Tench J, McMillan S, et al. Predicting patient visits to an urgent care clinic using calendar variables. *Acad Emerg Med*. 2001;8:48-53.

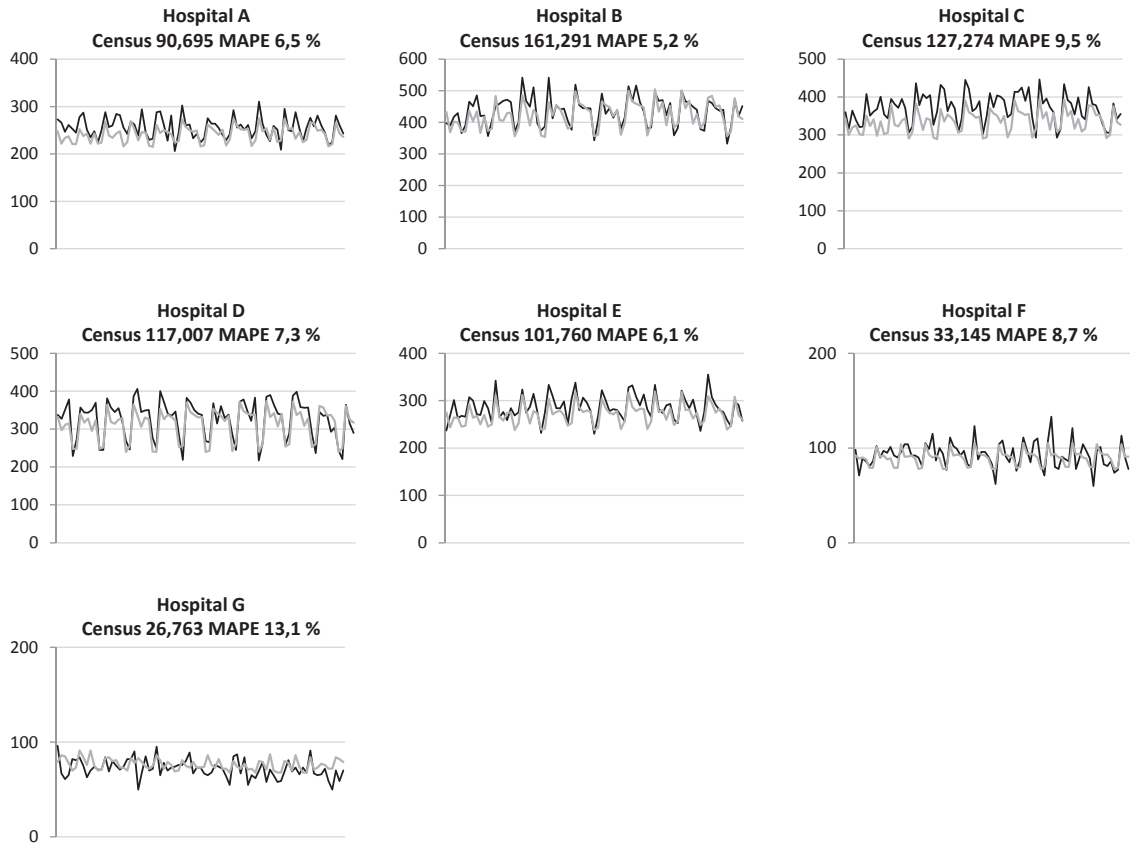

Figure E1. Predicted and measured values for all hospitals in Stockholm County during the test period, August 14, 2012, to October 30, 2012. Census is visits per the training set full year. The gray line represents prediction and the black line represents measured values.