A New Source of Data for Public Health Surveillance: Facebook *Likes*

Steven H. Gittelman
Elaine R. Trimarchi
Victor W. Lange

Objectives

Facebook *Likes* may be a source of digital data that can complement traditional public health surveillance systems and provide data at a local level. We explored the use of Facebook *Likes* as potential predictors of health outcomes and their behavioral determinants.

Abstract

We performed an exploratory quantitative analysis to examine the predictive qualities of Facebook *Likes* with regard to health outcomes (e.g., life expectancy, mortality, diabetes, and obesity) and lifestyle behaviors associated with chronic disease in 214 counties across the United States and 61 of the 67 counties in Florida. Data were obtained from both the 2010 and 2011 Behavioral Risk Factor Surveillance (BRFSS) and the National Vital Statistics Systems. Predictors were county-specific proportions of users with activity-, interest-, retail and shopping-related *Likes* selected for their relationship to health. Facebook *Likes* proved to be an effective predictor of all examined health outcomes and health behaviors. There was a persistent predictive benefit of Facebook *Likes* when added to socioeconomic status (SES) compared to the SES alone, though its magnitude varies widely. With the inclusion of Facebook *Likes,* effects range from an 11% improvement in variance explained when predicting obesity to a 353% improvement when predicting average duration since last routine checkup. Facebook Likes provide estimates for examined health outcomes and health behaviors that are comparable to those obtained from the BRFSS. Online sources may provide more reliable, timely, and cost effective county-level data than obtainable from traditional public health surveillance systems as well as serve as an adjunct to those systems.

Introduction

Big Data has the potential to revolutionize public health surveillance. The development of the Internet and the explosion of social media has provided many new opportunities for health surveillance. In 2013, Internet use among U.S. adults and adolescents aged 12 to 17 years has reached 80%-85%[1, 2] and 95%,[3] respectively, with the majority using wireless technologies to access the Internet, such as such as laptop computers, tablet computers, and cell phones or Smartphones.[4, 5] Moreover, the use of the Internet for personal health and participatory health research has exploded, largely due to the availability of online resources and healthcare information technology applications.[6-13] These online developments, plus a demand for more timely, widely available, and cost effective data, has led to new ways epidemiological data are collected, such as digital-disease surveillance, opt-in Internet panels, and Internet surveys.[13-30] For example, over the past two decades, Internet technology has been used to identify disease outbreaks, track the spread of infectious disease, monitor self-care practices among those with chronic conditions, and to assess, respond, and evaluate natural and manmade disasters at a population-level.[11, 13, 16, 17, 19, 20, 22, 27, 31-33] Use of these modern communication tools for public health surveillance has proven to be less costly and more timely than traditional population surveillance modes (e.g., mail surveys; telephone surveys; and face-to-face household surveys).

The Internet has spawned several sources of "Big Data," such as Facebook,[34] Twitter,[35] Instagram,[36] Tumblr,[37] Google,[38] and Amazon.[39] These online communication channels and market places provide a wealth of passively-collected data that may be mined for purposes of public health, such as socio-

demographic characteristics, lifestyle behaviors, and social and cultural constructs. Public health researchers need cost effective and readily available sources of health data at the local level and the Big Data revolution may provide a partial answer. Social networking sites, such as Facebook, have expanded to include over half of the US population,[40] allowing for digital data on Facebook users from virtually every area of the country. Moreover, researchers have demonstrated that these digital data sources can be used to predict otherwise unavailable information, such as socio-demographic characteristics among anonymous Internet users.[41-44] For example, Goel et al.[42] found no difference by demographic characteristics in the usage of social media and e-mail. However, the frequency with which individuals accessed the Web for news, healthcare, and research was a predictor of gender, race/ethnicity, and educational attainment, potentially providing useful targeting information based on ethnicity and income.[42] Integrating these big data sources into the practice of public health surveillance is vital to move the field of epidemiology into the 21st century as called for in the 2012 U.S. "Big Data" initiative.[24, 45]

Understanding how "Big Data" can be used to predict lifestyle behavior and health-related data is one step toward the use of these electronic data sources for epidemiologic needs.[42, 46] Facebook has been used by individuals and public health researchers for novel surveillance applications.[18, 43, 44, 47-50] Tong[44] reported on the use of Facebook as a surveillance tool among individuals involved in intimate partner break-ups. Chunara, et al.[18] used Facebook to examine the association between activity-related interests and sedentary-related interests and population obesity prevalence. These researchers found that populations with higher activity-related interests had a lower predicted prevalence of overweight and/or obesity. Facebook Likes are a means by which Facebook users can identify their own preferred Internet sites and interests. While Facebook Likes are not explicitly health-related, researchers have shown that when taken together, the 'network' of an individual's Likes are predictive of socio-demographics characteristics, health behaviors, obesity and health outcomes.[18, 43, 48, 50] Timian et al.[50] examined whether Facebook Likes for a hospital could be used to quickly and inexpensively evaluate two quality measures (i.e., 30-day mortality rates and patient recommendations). Facebook Likes have also been shown to be predictors of a variety of user attributes, such as intelligence, happiness, race, religious and political views, sexual orientation, and a spectrum of personality traits.[43] For example, *Likes* correctly predict homosexuality and heterosexuality, African American vs. White, and Democrat vs. Republican at levels above 85%. Researchers have proposed that Facebook Likes be used as a new behavioral measure in a fashion similar to traditional questionnaires.[43] The power of *Likes* is that they represent behavior.

In this study, we focus on harnessing the predictive power of Facebook *Likes* for the purpose of enhancing population health surveillance.
Towards this end, we view Facebook *Likes* as a class of "Big data" that may help us understand population health at a local level. To do this, the data we derive from Facebook *Likes* must be relevant to the health metrics we seek to address. *Likes* must predict life expectancy, the ultimate outcome of one's quality of health. Predicting intermediary causes of a shortened lifespan, such as obesity and diabetes, is also a worthwhile stepping stone to that goal. But in order to specifically target the risk factors associated with these conditions, *Likes* must also be able to predict the lifestyle behaviors that contribute to poor health outcomes. Given that risk factors and the associated disease are often clustered in populations geographically,[15, 51, 52] the ability to identify, monitor, and intervene at a population-level exists. If the Facebook characteristics of a region can predict physical activity, smoking, and self-care of chronic conditions (health maintenance), then a strong argument can be made in favor of the use of these data to target, monitor, and intervene on adverse lifestyle behaviors.

In this paper, we attempt to add to the scientific evidence-base on how "Big Data" might be used to complement traditional surveillance systems. We explored the use of Facebook *Likes* as potential predictors of health outcomes (e.g., morbidity, injury, disability, and mortality) and the behavioral determinants of poor health outcomes. Specifically, we hypothesized that: 1) Facebook *Likes* provide a means of characterizing communities; 2) Facebook *Likes* can be used as an indicator of chronic disease outcomes (obesity, diabetes, and heart disease); 3) Facebook *Likes* can be used as an indicator of mortality; and 4) Facebook *Likes* can be used as an indicator of adverse lifestyle behaviors that impact disease. If these hypotheses hold, then Facebook *Likes* can ultimately be used to enhance population health surveillance.

Methods

Data Sources

Data for the analysis were collected from a number of sources. Health outcome and risk behavior data were obtained from Behavioral Risk Factor Surveillance System (BRFSS). The BRFSS is an ongoing random-digit-dialed telephone survey operated by state health agencies with assistance from the Centers for Disease Control and Prevention (CDC). The surveillance system collects data on many of the behaviors and conditions that place adults (aged ≥18 years) at risk for chronic disease, disability, and death. The large sample size of the 2011 BRFSS (n = 506,467) facilitated the calculation of reliable estimates for 224 counties with 500 or more respondents. County-level risk factor data were obtained from the 2011 Selected Metropolitan/Micropolitan Area Risk Trends (SMART) subset of the BRFSS Health outcomes data (i.e., life expectancy, mortality, and low birth weight) were collected from the National Vital Statistics System (NVSS) which provides population data on deaths and births in the United States.[53] , Given the comprehensiveness of vital records, these data represent as complete a body of information on these statistics as can be achieved. As such, they may be considered the most reliable estimates employed by this study.

Facebook *Likes* data were collected using the Facebook Advertising API[54] in February 2013, which aggregates the number of users who express interest in certain categories of items by zip code. These zip code data were aggregated to the county-level to allow for direct comparisons to the health data[1]. The data reflect the cumulative total of Facebook users' likes at the time they are drawn. Out of 127 categories, 40 were selected for the model from the 'super-categories' of activities, interests, and retail and shopping[2]. Super categories were selected for their theorized relationship to health. For example, "Interests" contains the "Health & Wellbeing" category, to which the relationship of health is self-explanatory. The "Activities" category was chosen because it included "Outdoor Fitness & Activities," which seemed directly applicable to measures of physical activity, while "Retail & Shopping" was chosen due to its apparent linkage to socioeconomic status, a powerful driver of health outcomes.[55, 56] Other super-categories lacked these explicit links, though we acknowledge the possibility that potentially powerful indirect relationships may exist. Due to rounding performed automatically by the API that routinely led to overestimates, counties with fewer than 1,000 profiles overall were excluded from the analysis. Facebook *Likes* were scored as a percentage of completed profiles in an area. Finally, in order to reduce multicollinearity caused by variation in levels of Facebook usage by county, values were divided by the average percentage of *Likes* across all categories. The resulting variables can be characterized as a measure of popularity relative to that of other categories.[3]

Population data, such as average income, median age, and sex ratio, were collected using the 2010 U.S. Census[57] and broken into county aggregates. Supporting county-level statistics unrelated to health were collected using "USA Counties Information" provided by the Census Bureau.[58] Overall, 214 counties in the continental United States contained sufficient data for all variables in the analysis, while analysis of mortality data was possible in 2,879 counties.

Variables of Interest

Several sociodemographic, health outcome and risk factor variables were selected for analysis. These include income, age, education, employment, obesity, diabetes, physical activity, and smoking, as well as other measures such as general health status. A comprehensive listing, as well as the data source and assessment of each variable of interest are available in online appendix (see Appendix 1).

Data Analysis

[1] Zip codes crossing county borders were assigned to the county containing the largest geographic share

[2] The exact method for determining these categories has not been reported by Facebook.

[3] Though the individual variables resulting from this transformation were sometimes entirely uncorrelated with the originals, estimates using the raw and transformed variables correlated at R=0.9. Thus, we conclude that the results of the proceeding analyses are not an artifact of this transformation.

First, we used principal components analysis to reduce the 40 selected Facebook Likes categories to a more parsimonious set of factors that described the variation in these categories. We then used these factors in ordinary least squares (OLS) regression to determine whether Facebook *Likes* could predict life expectancy. Finally, we used these Facebook factors to predict other variables, beginning with the incidence of the diseases of diabetes and obesity and continuing on to predict a series of health-related behaviors.

Results

The first stage in the analysis was to establish that health outcomes could indeed be determined by Facebook *Likes*. Through principal components analysis, the forty categories were reduced to nine factors[4] (varimax rotation). Due to the complex structure of *Likes* contributing to these factors, we have resisted the urge to attempt to describe their meaning. Instead, each is numbered in accordance with the amount of variance it explains. The full matrix of loadings for the analysis can be found in Appendix 2.

In order to test our hypothesis that Facebook *Likes* can be used to predict mortality on their own, we used OLS regression. We used our nine Facebook factors to predict life expectancy, with no other controls included in this initial model. The results, as shown in the Facebook Only column of Table 1, were quite strong (model $R^2 = .69$). Despite this relationship, Facebook only has value insofar as it provides predictive value beyond that of reliable data that is already available through the census or other means. Regression results for an OLS model predicting life expectancy with demographic information on average age and socioeconomic status (as represented by average household income, unemployment rate and percentage with bachelor's degree) are shown in the socioeconomic status (SES) only column of Table 1. There is a very strong relationship to be found there as well, although it is less strong than for Facebook factors alone. Finally, the two groups of variables are combined in the last column of Table 1, indicating that while a great deal of the variance in life expectancy is shared by both the Facebook and SES variables, the addition of Facebook improves the model fit above and beyond readily available socioeconomic measures. The resulting R^2=0.80 also indicates that a considerable amount of the variation in county-level life expectancy can be explained by SES factors and Facebook likes.

Table 2 summarizes regressions run across an array of health variables and indicates the percent improvement in variance explained by the inclusion of Facebook *Likes* when added to SES compared to the SES alone. There are two conclusions we can draw from this model. First, Facebook *Likes* do prove to be an effective identifier of all tested disease outcomes. Second, there is a persistent benefit of Facebook *Likes* above and beyond that contributed by SES, though its magnitude varies widely.

Our next hypothesis stated that Facebook *Likes*, as a measure of personality or behavior, should be able to determine the behaviors that drive health outcomes. The results in Table 2 clearly show that the Facebook *Likes* factors had a sizeable impact in the predictive models of all tested health-related behaviors, and in some cases such as health insurance and exercise, the total model fit was quite strong.

Predicting Health Conditions

We have established the need for better estimates of health in small communities where survey data is insufficient. We believe a statistical model can be used for the purpose that incorporates Facebook *Likes*, but it is not necessary that Facebook *Likes* be the dominant force in the model. In our view, any variable that is available and reliable at a county level should be included in predictive models, regardless of the direction of its relationship with the measure in question. A number of the health measures used as dependent variables previously are extremely reliable non-survey statistics, and can incrementally increase model fit beyond what Facebook *Likes* and SES can do on their own.

[4] These were identified through examination of the scree plot that explained 85% of the variance.

Attempting to apply predictions from the 2011 SMART data creates a problem. Though predictions correlate well with actual levels in non-SMART data, mean levels are consistently upwardly biased. We hypothesize that the selection method that leads counties to be weighted according to the SMART program creates a non-representative sample with better levels of general health than we see in the United States in general, particularly in areas that are more rural. As an alternative without such problematic selection issues, we have limited our predictive model to 2010 Florida data. Florida collects over 500 interviews in 61 of its 67 counties every three years, leading to a dataset that has neither sample size shortages nor selection biases.

Using data exclusively from one state creates its own problems for a predictive model. Though the integrity of the data is very good, there is no easy way to correct for the various cultural differences between Florida and other states. Attempting to apply Florida-based models to the full set of SMART counties results in only fair level of correlation (R =0.63). Though it indicates that relationships exist, this is not a sufficient level of accuracy upon which to base policy decisions. Instead, we have limited our analysis to Florida in order to demonstrate the level of accuracy we feel can be achieved at a national level once a somewhat more representative selection of county-level data is made available.

The results of a predictive model are shown in Table 3. These are the averages of a 10-fold cross validation, where ~6 counties were randomly excluded and predicted with the remaining counties in each iteration. The inclusion of vitality statistics reduces but does not eliminate the contribution of Facebook *Likes* to the model. Although we would expect demographics and vitality statistics to be very effective at predicting "healthy" versus "unhealthy" communities, we believe that the additional data provided by Facebook *Likes* should help to clarify the finer distinctions between communities with similar general levels of health.[5]

Figures 1 & 2 show a graphical comparison of estimates versus source data in Florida, where nearly all counties were sufficiently sampled for reliable estimates. These maps are dynamically shaded from light to dark in accordance with the level of obesity. As should be apparent visually, the fit is generally good – 90% of errors in the model fall inside of ±2.1% (0.4 standard deviations) from CDC estimated values. The same process is repeated for general health in Figure 2.

Discussion

When we first undertook this research plan, it was our expectation that the larger part of the measurement error that would impact our results would come through the imprecise categorization and geographic aggregation of the Facebook data. But while there are some exceptions, the consistency and strength of fit we have found seem manifest. Our models do extremely well in predicting levels of health variables across counties where data are plentiful and often diverge from BRFSS estimates where they are not. This suggests the possibility that data imputed from Facebook and vital statistics may provide a more accurate picture in small counties than attempting to aggregate improperly balanced data across several years.

Thus, we argue that Facebook can serve an intermediary role in augmenting sparse data at a community level. We have shown that it can do so already, but additional health survey data, especially in less extensively measured regions (e.g., rural), could only help. Ensuring that communities of all types are represented in sufficient number when estimating the model is a necessary step in avoiding the risk of systematic error in its predictions.

The ultimate goal of our analysis of Facebook *Likes* is to establish the potential contribution of "Big Data" to research that directly impacts government spending and public policy, and, most importantly, contributes to improved population health. At a fraction of the cost of traditional research, data that might seem on its face to have little to do with health can predict life expectancy and epidemic-level health problems such as diabetes and obesity. With the need to augment

---

~~en predictions of two diseases dropped from 0.94 to 0.85 with the addition of Facebook likes (Z = 2.6, p<.05), which supports this theory.~~

traditional public health surveillance systems with readily available, cost effective, and geographically-relevant health data, the use of "big epidemiologic data" comes at just the right time.

Limitations

The nature of the Facebook data source prevents it from being a useful tool in several situations. In the case of very small counties (about 9%) and in any smaller geographic areas, rounding error becomes so great that estimates cannot be reliably used, even though they may be provided by Facebook. Facebook profiles are untested as a tool for tracking the prevalence of infectious diseases. They are better suited to predicting endemic and ongoing conditions that are unlikely to fluctuate over the course of short periods of time.

Conclusion

The relationships examined here demonstrate convincingly that social media can be used as an indicator of local conditions, even those that have little relationship to the activity that takes place on Facebook. As we predicted, significant relationships that extend beyond the predictive power of local demographics exist between an area's aggregate Facebook behavior and life expectancy, the incidence of diseases, and of health-related behaviors that very well may lead to those diseases.

We have also indicated the severe shortage of health data that is available and the great majority of American counties. While even Facebook data may not reach into every corner of the United States, it seems an effective enough tool to augment the existing county-level data in the majority of counties. With demand for local health data growing, such tools seem far more cost-effective than an increase in survey surveillance, regardless of the mode through which it might be conducted.

Whether this data ultimately comes from Facebook or not is of little importance. The online landscape may change, and it may provide a different source of data that proves more viable in the future. So long as the source reflects people's activities in daily life, the same relationships should hold. Even if Facebook does prove to endure as a social institution, however, there is still room for a great deal of improvement on the models presented here. With cooperation from the social media outlets themselves, we may be able to obtain better estimates in categories that align better with our needs. In the end, our data may not suffer as a result of the rising costs of research. Instead, exploring newly opened avenues of data collection online could lead to more reliable, timely, and cost effective data than ever.

REFERENCES

1. Centers for Disease Control and Prevention. Behavioral Risk Factor Surveillance System. Atlanta, GA: U.S. Department of Health and Human Services; 2013.
2. PewInternet. Trend data (adults). Washington, DC: PewResearchCenter; 2013.
3. PewInternet. Trend data (teens). Washington, DC: PewResearchCenter; 2013.
4. PewInternet. PewInternet. Pew Internet & American Life Project. Trend Data (Teens): Teen Internet Access Demograpics. *PewResearchCenter*. Available at: http://www.pewinternet.org/Static-Pages/Trend-Data-(Teens)/Whos-Online.aspx. Accessed July 17, 2013.
5. PewInternet. PewInternet. Pew Internet & American Life Project. Trend Data (Adults): Adult gadget ownership over time. *PewResearchCenter*. Available at: http://www.pewinternet.org/Static-Pages/Trend-Data-(Adults)/Device-Ownership.aspx. Accessed July 17, 2013.
6. Hand E. Citizen science: People power. *Nature.* Aug 5 2010;466(7307):685-687.
7. Brownstein CA, Brownstein JS, Williams DS, 3rd, Wicks P, Heywood JA. The power of social networking in medicine. *Nat Biotechnol.* Oct 2009;27(10):888-890.
8. Boicey C. Innovations in social media: the MappyHealth experience. *Nurs Manage.* Mar 2013;44(3):10-11.
9. Yu B, Willis M, Sun P, Wang J. Crowdsourcing participatory evaluation of medical pictograms using Amazon mechanical turk. *J Med Internet Res.* 2013;15(6):e108.
10. Chawla NV, Davis DA. Bringing Big Data to Personalized Healthcare: A Patient-Centered Framework. *J Gen Intern Med.* Jun 25 2013.
11. Rogstadius J, Vukovic M, Teixeira C, Kostakos V, Karapanos E, Laredo J. CrisisTracker: Crowdsourced social media curation for disaster awareness. *IBM J R&D.* 2013;57(5).
12. Eysenbach G. Medicine 2.0: social networking, collaboration, participation, apomediation, and openness. *J Med Internet Res.* 2008;10(3):e22.
13. Weitzman ER, Kelemen S, Mandl KD. Surveillance of an online social network to assess population-level diabetes health status and healthcare quality. *Online J Public Health Inform.* 2011;3(3):3797.
14. Brownstein JS, Freifeld CC, Madoff LC. Digital disease detection--harnessing the Web for public health surveillance. *N Engl J Med.* May 21 2009;360(21):2153-2155, 2157.
15. Salathe M, Bengtsson L, Bodnar TJ, et al. Digital epidemiology. *PLoS Comput Biol.* 2012;8(7):e1002616.
16. Morse SS. Public health surveillance and infectious disease detection. *Biosecur Bioterror.* Mar 2012;10(1):6-16.
17. Chunara R, Andrews JR, Brownstein JS. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *Am J Trop Med Hyg.* Jan 2012;86(1):39-45.
18. Chunara R, Bouton L, Ayers JW, Brownstein JS. Assessing the online social environment for surveillance of obesity prevalence. *PLoS One.* 2013;8(4):e61373.
19. Ayers JW, Althouse BM, Allem JP, et al. Novel surveillance of psychological distress during the great recession. *J Affect Disord.* Dec 15 2012;142(1-3):323-330.
20. Schmidt CW. Trending now: using social media to predict and track disease outbreaks. *Environ Health Perspect.* Jan 2012;120(1):A30-33.
21. Waggoner MR. Parsing the peanut panic: The social life of a contested food allergy epidemic. *Soc Sci Med.* 2013;90:49-55.
22. Chary M, Genes N, McKenzie A, Manini AF. Leveraging social networks for toxicovigilance. *J Med Toxicol.* Jun 2013;9(2):184-191.
23. Lauer MS. Time for a creative transformation of epidemiology in the United States. *JAMA.* Nov 7 2012;308(17):1804-1805.
24. Khoury MJ, Lam TK, Ioannidis JP, et al. Transforming epidemiology for 21st century medicine and public health. *Cancer Epidemiol Biomarkers Prev.* Apr 2013;22(4):508-516.
25. Crawford CAG, Okoro CA, Akcin HM, Dhingra S. An experimental study using opt-in Internet panel surveys for behavioral health surveillance. *Online J Public Health Inform.* 2012;5(1):e24.

26. Liu H, Cella D, Gershon R, et al. Representativeness of the Patient-Reported Outcomes Measurement Information System Internet panel. *J Clinical Epidemiology.* 2010;63:1169-1178.

27. Minniear TD, McIntosh EB, Alexander N, Weidle PJ, Fulton J. Using electronic surveys to gather information on physician practices during a response to a local epidemic-Rhode Island, 2011. *Ann Epidemiol.* 2013:1-3.

28. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. *J Med Internet Res.* 2009;11(1):e11.

29. Lyon A, Nunn M, Grossel G, Burgman M. Comparison of web-based biosecurity intelligence systems: BioCaster, EpiSPIDER and HealthMap. *Transbound Emerg Dis.* Jun 2011;59(3):223-232.

30. Hay SI, George DB, Moyes CL, Brownstein JS. Big data opportunities for global infectious disease surveillance. *PLoS Medicine.* 2013;10(4):e1001413.

31. Hingle M, Yoon D, Fowler J, et al. Collection and visualization of dietary behavior and reasons for eating using twitter. *J Med Internet Res.* 2013;15(6):e125.

32. Yoon S, Elhadad N, Bakken S. A practical approach for content mining of tweets. *Am J Prev Med.* Jul 2013;45(1):122-129.

33. Merolli M, Gray K, Martin-Sanchez F. Health outcomes and related effects of using social media in chronic disease management: A literature review and analysis of affordances. *J Biomed Inform.* May 20 2013.

34. Saverin E, Zuckerberg M, Moskovitz D, Hughes C. Facebook, Inc. Menlo Park, CA; 2013.

35. Dorsey J, Stone B, Williams E. Twitter, Inc. San Francisco, CA; 2013.

36. Systrom K, Krieger M. Instagram; 2013.

37. Arment M, Karp D. tumblr. New York, NY; 2013.

38. Page L, Brin S. Google. Mountain View, CA; 2013.

39. Bezos JP. Amazon.com. Seattle, WA: Amazon; 2013.

40. Smith C. (July 2013) How many people use the top social media, apps, & services? *DMR: Digital Marketing Ramblings...* 2013.

41. Murray D, Durrell K. Inferring demographic attributes of anonymous Internet users. In: Masand B, Spiliopoulou M, eds. *Web usage analysis and demographic profiling.* Vol 1836. Quebec, Canada: Springer Berlin Heidelberg; 2000:7-20.

42. Goel S, Hofman JM, Sirer MI. Who does what on the Web: A large-scale study of browsing behavior. Paper presented at: ICWSM, 2012; Toronto, Canada.

43. Kosinski M, Stillwell D, Graepel T. Private traits and attributes are predictable from digital records of human behavior. *Proc Natl Acad Sci U S A.* Apr 9 2013;110(15):5802-5805.

44. Tong ST. Facebook use during relationship termination: Uncertainty reduction and surveillance. *Cyberpsychol Behav Soc Netw.* 2013 Jun 20 2013:1-6.

45. Mervis J. U.S. science policy. Agencies rally to tackle big data. *Science.* Apr 6 2012;336(6077):22.

46. Bond RM, Fariss CJ, Jones JJ, et al. A 61-million-person experiment in social influence and political mobilization. *Nature.* Sep 13 2012;489(7415):295-298.

47. Chang A, Anderson EE, Turner HT, Shoham D, Hou SH, Grams M. Identifying potential kidney donors using social networking web sites. *Clin Transplant.* 2013;27(3):e320-326.

48. Lewis K, Kaufman J, Gonzalez M, Wimmer A, Christakis N. Tastes, ties, and time: A new social network datset using Facebook.com. *Social Networks.* 2008;30:330-342.

49. Jernigan C, Mistree BFT. Gaydar: Facebook friendships expose sexual orientation. *First Monday.* 2009;14(10).

50. Timian A, Rupcic S, Kachnowski S, Luisi P. Do patients "like" good care? Measuring hospital quality via Facebook. *Am J Med Qual.* 2013.

51. Mobley LR, Finkelstein EA, Khavjou OA, Will JC. Spatial analysis of body mass index and smoking behavior among WISEWOMAN participants. *J Womens Health (Larchmt).* 2004;13(5):519-528.

52. Schuit AJ, van Loon AJM, Tijhuis M, Ocke M. Clustering of lifestyle risk factors in a general adult population. *Prev Med.* 2002;35(3):219-224.

53. Centers for Disease Control and Prevention. National Vital Statistics System. *Centers for Disease Control and Prevention*. July 22, 2013. Available at: http://www.cdc.gov/nchs/nvss.htm. Accessed July 25, 2013.

54. Facebook. Facebook Ads API. *Facebook*. Available at: http://www.facebook.com/help/489212331104318. Accessed July 25, 2013.

55. Adler NE, Boyce WT, Chesney MA, Folkman S, Syme SL. Socioeconomic inequalities in health. No easy solution. *Jama.* Jun 23-30 1993;269(24):3140-3145.

56. Murray CJ, Abraham J, Ali MK, et al. The State of US Health, 1990-2010: Burden of Diseases, Injuries, and Risk Factors. *JAMA.* Jul 10 2013.

57. U.S. Census Bureau. United States' Census 2010: It's in our hands. *U.S. Census Bureau*. Available at: http://www.census.gov/2010census/. Accessed July 25, 2013.

58. U.S. Census Bureau. USA counties information. *U.S. Census Bureau*. Available at: http://www.census.gov/support/USACdata.html. Accessed July 25, 2013.

59. Gerbing D, Anderson J. On the meaning of within-factor correlated measurement errors. *J Consumer Research.* 1984;11(1):572-580.

TABLE 1—Ordinary Least Squares Regression Coefficients for Life Expectancy (All Independent Variables are Standardized)

	Facebook Only	SES Only	Facebook & SES
FB Factor 1	-0.4586**	-	-0.0302
FB Factor 2	1.2112**	-	0.8461**
FB Factor 3	-0.9336**	-	-0.3356**
FB Factor 4	0.4112**	-	0.5662**
FB Factor 5	0.4947**	-	0.3774**
FB Factor 6	0.1934**	-	-0.0411
FB Factor 7	-0.0511**	-	-0.0713**
FB Factor 8	0.2269**	-	0.1337**
FB Factor 9	-0.1147**	-	-0.0085
Age	-	0.3268**	0.0330
Income	-	0.8257**	0.7105**
Education	-	0.7158**	0.4419**
Unemployment	-	-0.3074**	-0.1084**
Constant	77.1254**	77.1254**	77.1254**
R^2	0.69	0.56	0.80

Note. FB = Facebook; SES = Socioeconomic status.
*P<.05; **P<.01.

TABLE 2—Facebook *Likes*' Impact on Model Fit

Dependent Variable	Source	SES (R^2)	SES + Facebook (R^2)	% Improvement with Facebook
Life Expectancy	NVSS	0.57	0.8	40%
Mortality	NVSS	0.43	0.63	47%
Low Birthweight	NVSS	0.17	0.57	235%
Obesity	BRFSS	0.56	0.62	11%
Diabetes	BRFSS	0.38	0.54	42%
Heart Attack	BRFSS	0.36	0.43	19%
Stroke	BRFSS	0.24	0.35	46%
Exercise	BRFSS	0.40	0.66	65%
Insured	BRFSS	0.19	0.55	189%
Fair/Poor Health	BRFSS	0.20	0.55	175%
Smoker	BRFSS	0.41	0.54	32%
Last Checkup	BRFSS	0.15	0.68	353%
Declined Treatment	BRFSS	0.16	0.48	200%

Note. %, percent; BRFSS, Behavioral Risk Factor Surveillance System; NVSS, National Vital Statistics System; SES, Socioeconomic status.

TABLE 3─ Ordinary Least Squares Regression Results for Prediction of Last Checkup[a]

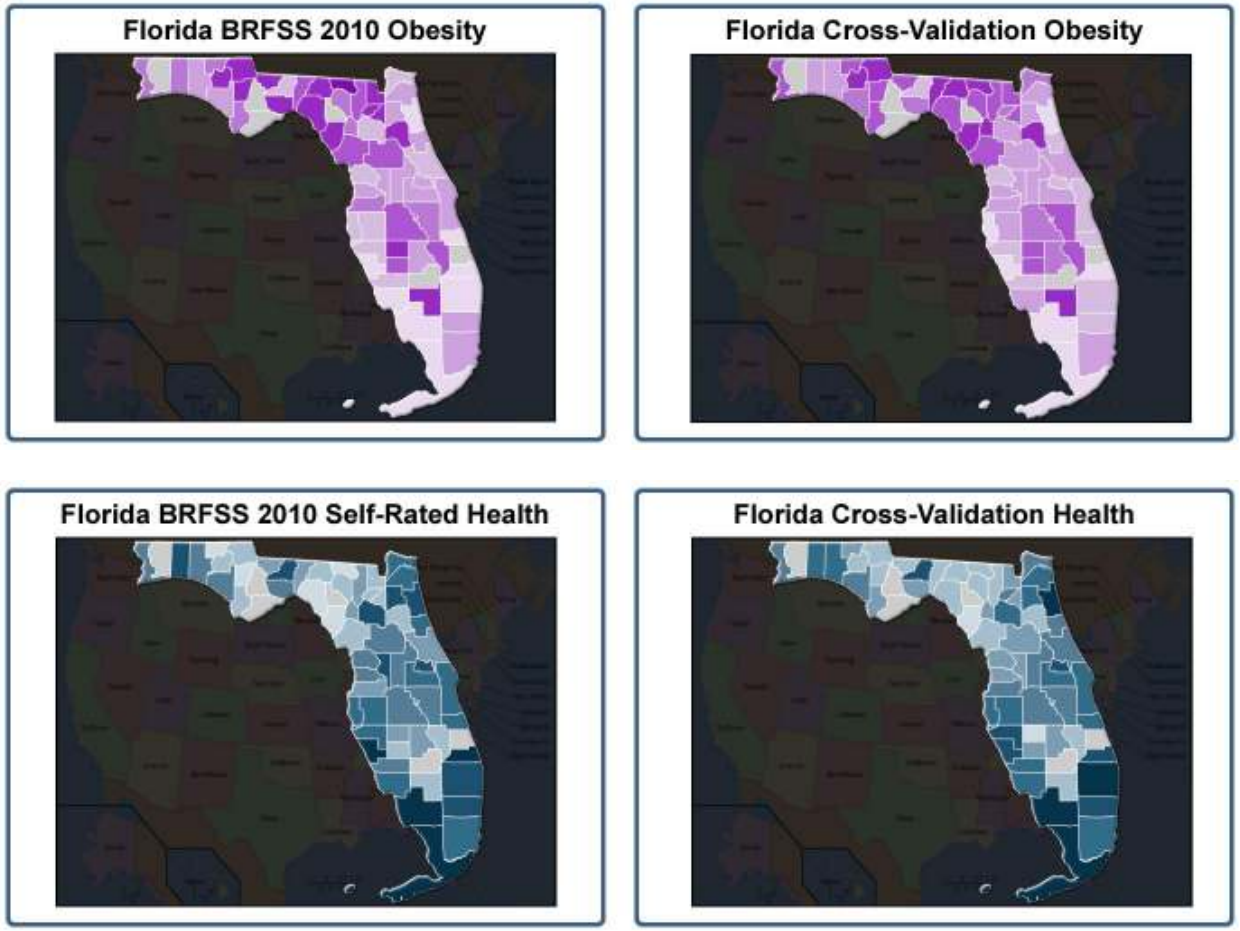
Variables	β	SE
FB Factor 1	0.032115**	0.009134
FB Factor 2	-0.01154	0.008358
FB Factor 3	0.018978*	0.009873
FB Factor 4	0.014612*	0.00759
FB Factor 5	-0.00444	0.007872
FB Factor 6	-0.01474*	0.008471
FB Factor 7	-0.00757	0.022947
FB Factor 8	-0.00026	0.010791
FB Factor 9	0.010992*	0.006287
Income	0.00219	0.010107
Age	-0.0122**	0.003354
Education	-0.01713*	0.009497
Unemployment	-0.00703	0.007161
Rural/Urban Scale	-0.00579	0.006783
Life Expectancy	0.00278	0.010803
Mortality	-0.00185	0.006593
% Underweight Births	0.019214*	0.00836
Constant	0.307739**	.020279
$R^2 = 0.87$		

Note. FB, Facebook.
[a]All independent variables standardized.
<<<Please distinguish between regression parameter estimates and standardized regression parameter estimates in the text and tables by:

1. changing all beta (b) symbols to b (for unstandardized regression parameter estimates) or B (for standardized regression parameter estimates); and
2. replacing all text or symbolic references to b in the manuscript and tables to language referencing b (parameter estimates) or B (standardized parameter estimates), as appropriate.

Beta (b), and other Greek symbols, should only be used in the text when describing the equations or parameters being estimated, never in reference to the results based on sample data.>>>

FIGURES 1& 2- Actual Statistics Compared with Predicted Values for Obesity and Self-Rated Health <<<Lower panel portrays fair/poor self-rated health; not diabetes>>>, 2010 BRFSS[a]



Note. BRFSS, Behavioral Risk Factor Surveillance System.
[a]Darker colors represent higher levels. Light gray indicates missing data.

Appendix 1: Variable Descriptions (Table missing heart attack and stroke)

Control Variables	Source	Question Wording or Description
Average Household Income	2010 Census	Mark the "Yes" box for each income source received during 2009 to a maximum of $999,999.
Median Age	2010 Census	What is this person's age and what is this person's date of birth?
Percent with bachelors degree	2010 Census	What is the highest degree or level of school this person has COMPLETED?
% Unemployed	Bureau of Labor Statistics (2010)	% in Labor Force without a job
Life Expectancy	National Vital Statistics System (2009)	Average age of death in a county
Adjusted Mortality	National Vital Statistics System (2009)	Age-Adjusted death rates
% of Underweight Births	National Vital Statistics System (2009)	% of babies born underweight
Obesity	BRFSS 2011 (SMART)	Body mass index > 30 based on self-reported height and weight
Diabetes	BRFSS 2011 (SMART)	Have you ever been told by a doctor that you have diabetes? (Diabetes caused by pregnancy excluded)
Physically Inactive	BRFSS 2011 (SMART)	No to: During the past month, other than your regular job, did you participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise?
Uninsured	BRFSS 2011 (SMART)	No to: Do you currently have health insurance?
Fair/Poor General Health	BRFSS 2011 (SMART)	In general, would you say your health is Excellent, Very Good, Good, Fair or Poor?
Smokes Every Day	BRFSS 2011 (SMART)	(To those who have smoked 100 cigarettes) Do you now smoke cigarettes every day, some days, or not at all?
Last Checkup	BRFSS 2011 (SMART)	About how long has it been since you last visited a doctor for a routine checkup?
Cost Barrier to Needed Healthcare	BRFSS 2011 (SMART)	Was there a time in the past 12 months when you needed to see a doctor but could not because of cost?
Heart Attack	BRFSS 2011 (SMART)	Has a doctor, nurse, or other health professional ever told you that you had a heart attack, also called a myocardial infarction?

Stroke	BRFSS 2011 (SMART)	Has a doctor, nurse, or other health professional ever told you that you had a stroke?

Appendix 2: Ancillary Tables

Table Ia & Ib: Rotated (Orthogonal Varimax) Factors and Loadings

	Eigenvalue	Difference	Proportion	Cumulative %
Factor1	8.15809	3.39113	0.2205	0.2205
Factor2	4.76697	0.80242	0.1288	0.3493
Factor3	3.96455	0.48799	0.1072	0.4565
Factor4	3.47656	0.38719	0.094	0.5504
Factor5	3.08937	0.12598	0.0835	0.6339
Factor6	2.9339	0.43231	0.0801	0.714
Factor7	2.53109	0.98687	0.0684	0.7824
Factor8	1.54422	0.50612	0.0417	0.8242
Factor9	1.031	.	0.0281	0.8522

Variable	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
Auto Intenders	0.9617	-0.0206	0.0506	-0.0335	0.0001	-0.0725	-0.003
Automobiles	0.3698	0.177	0.4983	0.4453	0.2792	0.0582	0.0489
Beauty	0.537	-0.1706	0.5095	-0.2078	-0.2586	0.0561	0.1883
Beer/Wine/Spirits	0.2143	0.7312	0.0987	0.0406	0.0798	-0.1622	0.3439
Charity	-0.1008	-0.8091	0.244	-0.038	-0.1489	0.1103	0.0665
Electronics	-0.4048	-0.0684	0.7577	-0.176	-0.1939	-0.1786	0.1409
Cooking	-0.0201	0.3403	-0.2026	0.1151	0.2414	0.7686	0.1273
Dancing	0.2056	-0.0511	0.0176	-0.0877	-0.0092	-0.1315	-0.9444
Do-It-Yourselfing	0.2259	0.2603	-0.0591	0.4813	0.3427	0.5401	0.1425
Teaching	0.8527	-0.0852	-0.046	-0.1278	-0.0368	-0.0866	-0.4094
Television	-0.6677	0.3328	0.1216	0.0645	0.2408	0.2605	0.2731
Environment	0.2074	0.3268	-0.3263	0.6258	0.4052	0.1468	0.0214
Planning	0.9613	-0.0424	0.0369	-0.0378	0.0068	-0.0732	-0.0121
Fashion	-0.6427	-0.1994	0.4196	-0.2189	-0.1012	-0.2475	0.2634
Fast Food	0.1112	-0.136	0.0866	-0.082	-0.9261	-0.1337	0.0497
Food & Dining	0.0419	0.0082	0.0021	-0.0072	0.0928	0.9079	0.0815
Frequent Casual Diners	-0.1064	-0.1554	0.0172	-0.0608	-0.9344	-0.1388	0.0831
Game Consoles	0.0494	0.2899	0.4897	-0.0439	0.0436	-0.203	0.0553
Social Gaming	-0.0266	0.0822	0.8227	0.2308	0.0454	-0.1446	0.1124
Gardening	0.8088	0.0324	-0.0031	0.3319	0.192	0.171	0.0439
Health & Wellness	0.014	0.4426	-0.4525	0.4023	0.4108	0.2875	0.0867
Home & Garden	0.1721	0.1869	0.0922	0.2671	0.2818	0.2568	-0.0433
Literature	-0.0926	0.1469	-0.1568	0.4665	0.3935	0.4739	0.0782
Luxury Items	0.3351	0.099	-0.1855	-0.7809	-0.2401	-0.0426	0.1214
News	-0.3251	0.3166	-0.3513	-0.1588	0.1599	0.2863	0.097
Outdoor Activities & Fitness	-0.0683	0.0995	-0.1691	0.1884	0.1308	-0.0445	-0.9235
Pets	0.1006	0.5345	0.0872	0.3108	0.1958	0.3842	0.1424
Cats	0.9165	0.1689	0.103	-0.0442	0.0054	0.1107	0.0205

Variable	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
Dogs	0.7732	0.4081	0.0819	0.1486	0.1385	0.0239	0.0569
Photo Uploads	0.8038	0.0517	0.0238	-0.1711	-0.2023	0.0203	0.0273
Photography	0.7768	0.1544	-0.3233	0.0586	0.1081	0.2214	0.065
Politics	-0.2038	-0.7978	-0.088	-0.359	-0.0424	-0.2274	0.1694
Conservative Politics	-0.06	-0.7883	-0.0923	0.1785	-0.1902	-0.3164	0.143
Liberal Politics	-0.0113	-0.2381	-0.0724	-0.8902	0.1221	-0.0034	0.0546
Nonpartisan Politics	-0.2404	0.8188	-0.0362	0.2682	0.1204	0.1603	0.0523
Pop Culture	-0.3853	0.2409	0.5176	0.0767	-0.0714	0.0621	0.3114
Travel	-0.2209	0.0504	-0.8341	-0.0663	0.0519	-0.1225	0.056

Variable	Factor8	Factor9	Uniqueness
Auto Intenders	0.0731	-0.0683	0.0557
Automobiles	-0.2184	0.1008	0.2437
Beauty	-0.2353	0.1445	0.198
Beer/Wine/Spirits	-0.0481	-0.2966	0.1668
Charity	-0.2328	0.1228	0.1662
Electronics	0.0683	0.1537	0.1088
Cooking	0.1536	-0.0495	0.1383
Dancing	0.0196	0.0245	0.0369
Do-It-Yourselfing	0.2333	-0.0669	0.1577
Teaching	-0.0632	-0.0638	0.0627
Television	-0.264	0.0027	0.1543
Environment	0.1359	0.0562	0.1443
Planning	0.0517	-0.0868	0.0555
Fashion	0.1254	-0.0359	0.1654
Fast Food	-0.1002	0.0468	0.0647
Food & Dining	0.0603	-0.0443	0.153
Frequent Casual Diners	-0.0871	-0.0549	0.0506
Game Consoles	0.2291	0.6379	0.1662
Social Gaming	0.1367	0.1547	0.1843
Gardening	0.1765	-0.1569	0.1109
Health & Wellness	0.0981	0.0649	0.1645
Home & Garden	0.7878	0.1349	0.0696
Literature	-0.0708	0.1615	0.311
Luxury Items	-0.1079	0.1557	0.1236
News	0.4123	0.0895	0.3506
Outdoor Activities & Fitness	0.0254	-0.045	0.0467
Pets	-0.1453	0.2468	0.3119
Cats	-0.0046	0.1111	0.0939
Dogs	-0.059	0.1952	0.1423
Photo Uploads	-0.0699	0.0941	0.2656
Photography	0.0952	0.1413	0.1708

Variable	Factor8	Factor9	Uniqueness
Politics	-0.0715	-0.2392	0.0409
Conservative Politics	-0.0303	-0.2351	0.1217
Liberal Politics	-0.0451	-0.0374	0.1242
Nonpartisan Politics	-0.0224	0.0484	0.1527
Pop Culture	-0.4353	0.1707	0.1952
Travel	0.0899	0.1453	0.1985