Online Supplement

Materials and Methods

Mapping Tweets to Counties

The method described in Schwartz et al. (2013a) was used to map language expressed on Twitter (tweets) to counties. This method relies on either the coordinates attached to a tweet (latitude, longitude) or the free-response "location" field for the Twitter user who posted the tweet to determine the tweet location. One percent of our tweets had coordinates. To map a pair of coordinates to a county, the point given by the coordinates was checked to see whether it was within the boundaries of a U.S. county. Other tweets were mapped to counties by the location text field. If the location field included city and state, we matched to the relevant county. For location fields with only city information, we could match counties if the name was unambiguous, defined as having a 90% likelihood of being one particular according to census population statistics (e.g., Chicago was unambiguously Chicago, Illinois, whereas Springfield could easily be Springfield in Pennsylvania, Virginia, or elsewhere). Large non-U.S. cities were also thrown out (e.g. London). This method favored fewer false positives (incorrect mappings) at the expense of mapping a more limited number of tweets. To access accuracy of this mapping process, human raters judged a sample of 100 tweets; 93% were true positives (correct mappings)**.** Approximately 16% of the tweets could be mapped to U.S. counties (about 148 million tweets).

Tokenization

Tokenization is the process of splitting sentences into words (also known as "tokens"). Typically, this involves identifying sequences of letters separated by spaces and disjoining punctuation where appropriate (e.g., "The C.D.C. reports heart disease rates aren't increasing."

gets separated into "The", "C.D.C.", "reports", "heart", "disease", "rates", "aren't", "increasing", and "."). We used a tokenizer designed for social media that accurately captures emoticons such as ":)" (a smile) or "<3" (a heart) as words (Schwartz et al., 2013b). At the county-level, the frequencies of every unique word were summed, giving the word use for the county. From there, the dictionary and topic language features were derived.

Topic Extraction

Topics contain lists of semantically-related words. Unlike dictionaries, they are derived automatically, using a well-established algorithm from computer science, Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). LDA is a Bayesian mixture model, which groups words together that often appear together (e.g., one topic included *infection*, *ear*, *doctor*, *sinus*, *meds*, *antibiotics*, *poor*, *medicine*). Topics allow consideration of unanticipated categories of words. We used 2000 topics (available at wwbp.org/data.html) and derived their probability of usage per county from the relative word frequencies, as described in Schwartz et al. (2013b). The prevalence of a word in a topic is defined as the frequency for which that word appeared in the topic during fit of the LDA model. In other words, it is the estimated frequency for which the word is used as a representative of the topic.

Predictive Models

Our cross-sectional predictive models were fit via ridge regression (Hoerl & Kennard, 1970), which uses a standard machine learning approach of penalizing variable weights to avoid over-fitting due to variable multicollinearity. A10-fold cross-validation approach was used to fit and test models. Specifically, all 1,347 counties were divided randomly into ten nearly-equal sized groups ("folds"); nine folds were used as the "training set" in order to fit the model, and the final fold was used to test the model. The ridge regression method includes a penalization

parameter (often called "alpha"), and we also used univariate feature selection, which includes a parameter automatically set by the algorithm by testing on a subset of the training data. Predictive accuracies (performance) were recorded as a Pearson *r* correlation between the predicted mortality rates and the Centers for Disease Control and Prevention (CDC) reported mortality rates (2010). The process was then repeated 10 times, such that a new fold became the test set each time, and the predictive accuracies were averaged across the 10 runs. Standard errors of the predictive accuracies were based on the accuracies across these 10 runs.

When using language features, we had many more independent variables (i.e., tens of thousands of language features) than we did units of analysis (counties). To avoid overfitting, we used univariate feature selection fed into Principal Component Analysis (PCA) (Hotelling, 1933; Martinsson, Rokhlin, Tygert, 2011*)* for each type of independent variable (i.e. running the word and phrase features separately from topics). In univariate feature selection for regression, we removed individual features that were not significantly correlated at a family-wise alpha of 60 with the mortality rates. PCA then reduced the number of dimensions to either 10% of its original size or half the number of counties – whichever was smaller. Both the significance level and dimensional reduction size were selected based on tests over the training sample. Such steps are common practice in the field of machine learning when dealing with large numbers of independent variables (Hastie, Tibshirani, Friedman, 2009). When creating a model based on non-language variables (i.e. the health and demographic values; at most 10 variables at a time), we entered the variables as independent variables into the linear ridge regression model without using univariate feature selection or dimensionality reduction, as these steps are unnecessary with simple conventional independent variables in a regression model.

References

American Community Survey (ACS). (2009). Selected social characteristics in the United States. Retrieved from http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_09_1YR_DP2&prodType=table

Behavioral Risk Factors Surveillance Survey. (2009-2010). Annual survey data. Centers for Disease Control and Prevention. Retrieved from www.cdc.gov/brfss/annual_data/annual_data.htm

Blei, D. M., Ng, A. Y., & Jordan, M.I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3,* 993-1022.

Centers for Disease Control and Prevention (CDC). (2010). Underlying cause of death 1999-2010. *CDC WONDER Online Database*. Retrieved from http://wonder.cdc.gov/ucd-icd10.html

County Health Rankings and Roadmaps. (2010). Rankings data. Retrieved from http://www.countyhealthrankings.org/rankings/data

Diabetes Public Health Research. (2010). Diagnosed diabetes prevalence. Centers for Disease Control and Prevention. Retrieved from http://www.cdc.gov/diabetes/atlas/countydata/County_EXCELstatelistDM.html

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*, 2nd ed. Springer. http://dx.doi.org/10.1007/978-0-387-84858-7

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12,* 55-67. http://dx.doi.org/10.1080/00401706.1970.10488634

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24,* 417. http://dx.doi.org/10.1037/h0071325

Institute for Health Metrics and Evaluation (IHME). (2009). United States hypertension estimates by county 2001-2009. Global Health Data Exchange website. Retrieved from http://ghdx.healthmetricsandevaluation.org/record/united-states-hypertension-estimates-county-2001-2009

Martinsson, P. G., Rokhlin, V., & Tygert, M. (2011). A randomized algorithm for the decomposition of matrices. *Applied and Computational Harmonic Analysis, 30,* 47-68. http://dx.doi.org/10.1016/j.acha.2010.02.003

National Health Examination and Nutrition Survey. (2010). Centers for Disease Control and Prevention. Retrieved from http://www.cdc.gov/nchs/nhanes.htm

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Lucas, R. E., Agrawal, M., …, & Ungar, L. H. (2013a). Characterizing Geographic Variation in Well-Being using Tweets. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM)*. Boston, MA.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., …, & Ungar, L. H. (2013b). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS ONE, 8,* e73791. http://dx.doi.org/10.1371/journal.pone.0073791