Future-Oriented Tweets Predict Lower County-Level HIV Prevalence in the United States

Molly Ireland, Andy Schwartz, Lyle Ungar, & Dolores Albarracin

Paper was Invited for Resubmission, which we did in May 2015

Abstract

Objective

Future orientation promotes health and well-being at the individual level. Computerized text analysis of a dataset encompassing billions of words used across the United States on Twitter tested whether community-level rates of future-oriented messages correlated with lower HIV rates and moderated the association between behavioral risk indicators and HIV.

Method

Over 150 million Tweets mapped to US counties were analyzed using two methods of text analysis. First, county-level HIV rates (cases per 100,000) were regressed on aggregate usage of future-oriented language (e.g., *will*, *gonna*). A second data-driven method regressed HIV rates on individual words and phrases.

Results

Results showed that counties with higher rates of future tense on Twitter had fewer HIV cases, independent of strong structural predictors of HIV such as population density. Future-oriented messages also appeared to buffer health risk: Sexually transmitted infection rates and references to risky behavior on Twitter were associated with higher HIV prevalence in all counties except those with high rates of future orientation. Data-driven analyses likewise showed that words and phrases referencing the future (e.g., *tomorrow*, *would be*) correlated with lower HIV prevalence.

Conclusion

Integrating big data approaches to text analysis and epidemiology with psychological theory may provide an inexpensive, real-time method of anticipating outbreaks of HIV and etiologically similar diseases.

Key words: language, Twitter, HIV, epidemiology, future orientation, risk

Future-Oriented Tweets Predict Lower County-Level HIV Prevalence in the United States

Clinical, social, and health psychology tend to agree that thinking about and planning for the future promotes health and well-being, a principle that has influenced research on public health, economics, and safety (Chen, 2013; Ebreo & Vining, 2001; Hoyle & Sherrill, 2006). Seeking long-term rewards and thinking about the future is also associated with lower impulsivity and addictive behavior as well as better overall health (Boyd & Zimbardo, 2005; Daugherty & Brase, 2010; MacKillop et al, 2011; Weller, Cook, Avsar, & Cox, 2008). More germane to the present research, people who are more future-oriented engage in less risky sexual behaviors and take a more proactive approach to reducing HIV risk from sex (Rothspan & Read, 1996). The increasing availability of big data, especially in the form of online language use, has given behavioral scientists the unprecedented opportunity to test this cross-cutting psychological principle in the crucible of real-life behavior. This research examined whether communities' future tense verb usage on Twitter relates to HIV prevalence across counties in the United States. Our approach used epidemiologic methods of HIV forecasting and computational linguistic tools (Pennebaker, Booth, & Francis, 2007; Schwartz et al., 2013) to test this tenet, with the broad aim of improving disease transmission models that predict where outbreaks will occur next.

The relation between thinking about the future and health is likely bidirectional. Although people may be more effective at reaching goals (including those related to health) when they think about and plan for the future, effective self-regulation and meeting goals should also allow individuals to disengage from prior goals and focus on the future. An analysis of Google search terms found that queries containing references to future years (i.e., "2013" in 2012) are more common in wealthier countries than their poorer counterparts (Preis, Moat, Stanley, & Bishop, 2012). These results mirror earlier findings that adolescents from families of higher socioeconomic status are more future-oriented than those from lower status families (Nurmi, 1987). Indeed, poorer individuals' tendency to think less about the future appears to explain underprivileged groups' reluctance to take part in preventive healthcare, such as cancer screenings (Whitaker et al., 2011). A future orientation is required for people to anticipate and monitor their health.

Importantly, future orientation is not synonymous with optimism or positivity (Zimbardo & Boyd, 1999). People tend to reap the benefits of thinking about the future whether their thoughts are positive or negative, and whether they think of themselves as approaching an ideal self or avoiding a feared self (Markus & Nurius, 1986). Additionally, when envisioning these possible selves, it is better to set practical and achievable goals than to set highly optimistic goals that may be impossible to reach (Oyserman, Bybee, & Terry, 2006). Thinking about possible selves motivates proactive behavior and improves self-regulation to the degree that a person's current and possible future selves are discrepant; in this way, those who currently are satisfied with their health or well-being can maintain preventive health behaviors, such as regular exercise, by thinking about feared, less healthy future selves (Hoyle & Sherrill, 2006; Markus & Nurius, 1986). Delayed gratification, another behavior associated with future orientation and better health, is also not necessarily bound to optimism: Forgoing current rewards in order to obtain long-term future goals can be framed positively (e.g., receive more marshmallows if you wait) or negatively (e.g., avoid drug use in order to postpone death; Mischel & Mischel, 1983; Green, Frye, & Myerson, 1994).

For these reasons, we predicted that thinking about the future in a broad sense – rather than thinking about the future in positive terms specifically – would be associated with decreased HIV rates and would attenuate HIV risk in vulnerable communities. In theory, communities that think and talk about the future more frequently will be healthier in several respects, including offering easier access to preventive healthcare (if community leaders are future oriented) and providing fewer temptations to engage in risky behavior (if individuals in that community navigate their ideal and feared future selves by regulating their present behavior; Hoyle & Sherrill, 2006).

Beyond socioeconomic structures that increase individuals' health risk, risky behaviors (such as drug use) increase communities' vulnerability to HIV transmission and other health problems. Analyzing natural language use[1] provides a simple, face-valid means of assessing the degree to which members of communities think about, and presumably engage in, risky behaviors. Tweeted references to sex and drug use, for example, positively correlate with counties' HIV rates, controlling for standard socioeconomic predictors of HIV (Young, Rivers, & Lewis, 2014). We, however, propose that the risks of living in a community where risky behavior is common may be attenuated in relatively future-focused communities. Put another way, the effects of risky behavior, such as substance abuse, will theoretically be strongest in communities that carry the additional risk of lacking forethought, or failing to think about the future. The degree to which linguistic references to risky behavior, such as drinking or drug use, reflect the actual risks that people take in those communities will likely be qualified by other aspects of individuals' language use, such as future orientation. For example, sex itself does not lead to or increase HIV risk unless it happens in a high-infectivity area with low frequency of prevention, testing, and treatment. Communities that talk about sex but do so in a way that suggests planning for the future may even be at less risk for future HIV outbreaks than those that avoid discussing sex – a prediction supported by many successful HIV-prevention interventions that focused on discussing barriers to safer sex with members of at-risk groups (Albarracin et al., 2005).

Past studies of future-orientation have relied primarily on either self-report measures or, in the case of language, single search queries to index the degree to which a person thinks about the future. With the rise of social media popularity, we increasingly have access to records of naturalistic behavior from the large and diverse group of individuals who spend part of their daily lives communicating online. Twitter is used by about 16 percent of respondents polled in the United States, a number that is consistent across genders, education levels, and age groups up to age 50 (Duggan & Brenner, 2012). Researchers so far have used Twitter to track the spread of disease (Lampos & Cristianini, 2010; Sadilek, Kautz, & Silenzio, 2012) and assess geographical variation in well-being (Schwartz et al., 2013).

The present study used both theory- and data-driven text analysis methods to measure future orientation in natural language used on Twitter, aggregated at the county level in the United States. We predicted that future-orientation will (1) be associated with fewer HIV cases and (2) buffer HIV risk in more vulnerable counties, indicated by higher prevalence of common sexually transmitted infections (STIs), including chlamydia, gonorrhea, and syphilis (Fleming & Wasserheit, 1999) and more frequent Twitter references to risky behavior (e.g., *bars*, *gambling*; see Young, Rivers, & Lewis, 2014).

Method

Our final collection of tweets mapped to US counties included over 150 million messages. Tweets were randomly sampled from those posted between June 2009 and March 2010 and comprised 10% of all tweets sent in that period. We excluded counties from which we had less than 10,000 commonly used words (*n* = 1,046), resulting in a dataset of 2,079 counties.

Geolocation. Tweets do not have county information directly attached, and mapping tweets to locations is not straightforward (Hecht, Hong, Suh, & Chi, 2011). For the approximately 2% of tweets that have geolocation coordinates, it is simply a matter of finding which county the coordinates reside within. However, for the rest of the data we used the free-response location field that accompanies a tweet. This field may contain a city/state pair or an individual city, but also often has unhelpful phrases (e.g., "behind you"). We used the set of rules described in Schwartz et al. (2013) to map location fields to counties. The locations fields were broken up into sequences of words (tokenization) and then matched to country names. Out of those messages either mentioning the country as the United States or not mentioning a country, we used the words preceding the country and attempt to match city and state names. We used city population information when only a city was matched to determine whether the city was 90% likely to be in any state; if so, we paired the city with its most likely state. Otherwise, the tweet was discarded. For example, if Springfield, Illinois has a population of approximately 117,000 and the sum of populations across all cities named Springfield is 187,000, then we would calculate the likelihood that "Springfield" is referring to Springfield, Illinois as 117,000 / 187,000 = 62.6 percent; thus, Springfield would not be mapped. We also excluded city names that also happened to match one of the 100 largest non-US cities. Of the total sample of tweets, 78.6 percent were discarded due to either a geolocation outside the U.S., lack of any geolocation data, or ambiguous geolocation (from 706 million tweets down to approximately 151 million).

As the geolocation process is designed to be highly precise (i.e., few false positives), it may disregard many tweets. Schwartz et al. (2013) found that 93% of a sample of 100 city-state pairs were mapped to the city and state correctly. The cities were then mapped onto their respective counties, as our data on HIV and STI prevalence were available at the county level.

Note that some tweets are generated automatically from bots. Many of these are filtered out during county-mapping. Whether tweets that get past this step still reflect differences between counties or simply add noise to our prediction problem is an open question that is outside the scope of this study.

Text Analysis. We analyzed word use in tweets using both theory- and data-driven methods. In the first set of analyses, we calculated the percentage of words in tweets from each county that referenced future tense (e.g., *will*, *should*), risky leisure activities (e.g., *clubbing*, *beer*), and comparatively safe leisure activities (e.g., *picnic*, *shopping*) using the Linguistic Inquiry and Word Count (LIWC; Pennebaker et al., 2007). LIWC compares all words in a given text against a number of internal dictionaries, or word lists. These categories include parts of speech (e.g., conjunctions, articles), psychological dimensions (e.g., positive emotion, inhibition), and topics (e.g., work, sex). LIWC has been extensively validated and is among the most widely used computerized text analysis tools in the behavioral sciences (for a review, see Tausczik & Pennebaker, 2010).

Future tense is an existing category in LIWC2007. Risky and safe leisure activities were adapted from LIWC's leisure category. We first generated additional leisure words that might relate to risk, including a more complete list of recreational drugs and alcohol than the original dictionary. Two trained judges then rated each word in the resulting dictionary as either safe or risky, and disagreements were resolved through discussion. To avoid false positives, ratings erred on the side of labeling a word as safe. For example, although coffee and shopping can be considered, respectively, an addictive drug and an impulsive behavior, each word was considered “safe” for the purposes of this study. Risky language was calculated as the difference between the standardized LIWC output for risky and safe leisure activities (risky *z*-score – safe *z*-score). For additional examples of each category, see Table 1.

LIWC counts the number of words in a text that match a particular dictionary and then divides that number by the total words in the text to produce percentages. As the size of our Twitter dataset prevented us from manually inspecting every tweet for data validation, one or a few high-frequency words may produce significant results for an entire category (Back, Kufner, & Eglaf, 2011). For example, if a majority of words in a category are negatively correlated with HIV but one very high-frequency word shows the opposite pattern, the entire category could erroneously appear to be positively correlated with HIV. To avoid such misleading results for the future category, which is made up of a small number of high-frequency function words that should each equally reflect future orientation (e.g., *will* and *could*), we used a more conservative word count method in which each dictionary word's frequency is log transformed before being summed to calculate the output for each word category. This strategy ensures that each word in a dictionary contributes relatively equally to its category's overall score. The risky and safe leisure dictionaries, in contrast, contain a large number of low-frequency content words that vary widely in terms of how well they represent the category of risky behavior and how frequently they are used (e.g., the word *drunk* may be both riskier and more common than *blackjack*). Thus, we expect some of the risky and safe dictionary words to contribute to HIV risk differently and do not believe it would be theoretically useful to suppress those differences.

Second, to allow for the possibility of finding unpredicted patterns in the data, we used a bottom-up, data-driven approach termed Differential Language Analysis (DLA). Using DLA, words and phrases were extracted from tweets using a social media tokenizer, and topics (groups of semantically-related words) were identified. The topics were previously identified, in the context of social media, by running a Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003) process on a sample of 14 million Facebook posts (Schwartz et al., 2013). DLA then produces lists of individual words and phrases that most strongly correlate with a given outcome. These lists are visualized as word clouds that represent the correlation strength and overall frequency for the most highly correlated words and phrases.

Results

Results are organized into two main sections: (1) hypothesis tests that relied on the top-down method of comparing words in Tweets against LIWC's internal dictionaries, and (2) the exploratory, data-driven DLA method that considered words and phrases individually as predictors of HIV prevalence (i.e., diagnoses per 100,000 people per county, as reported by the CDC in 2010; AIDSVu, 2013).

Future Dictionary

The first set of analyses regressed HIV prevalence on future tense and structural control variables in multi-level Poisson regression models, including random slopes and intercepts to account for variation in beta-weights and mean HIV prevalence between states. All predictors were *z*-scored. Each model controlled for the three best structural correlates of county-level HIV prevalence: percentage of the population that identifies as Black, percentage of the population that is foreign born, and population density in the county (Prejean et al., 2011). Counties that had Twitter data but were missing data for HIV rates and each control variable were excluded from analysis ($n = 278$).

Because our ultimate aim is to forecast HIV outbreaks, we used out-of-sample cross-validation prediction to evaluate the associations between language use and HIV prevalence. That is, we ran two sets of analyses for each predicted main effect, each focusing on different subsets of the sample. Such an evaluation avoids overfitting – the situation where a model's coefficients reflect noise in the fit data rather than an association that generalizes to unseen data (Picard & Cook, 1984). The first set of analyses used data from all but 400 counties, approximately three-fourths of the counties (*n* = 1238) for which data on HIV prevalence were available, that were randomly selected using a random sequence generator (random.org). The second set of analyses ran the same models, fit to the first set of counties, over the remaining 400 counties to determine whether the features that initially correlated with HIV risk would generalize to the smaller subsample. Results are reported side by side. To conserve power, moderator analyses were run on the entire sample.

For both the initial subsample and the remaining counties (displayed side-by-side), HIV prevalence is lower in counties with higher rates of future tense (*B* = -0.27/-0.48, *SE* = .05/.09, 95% CI = [-0.18, -0.36]/ [-0.30, -0.66], *z* = -5.88/-5.21, both *p*s <.001).[2] In separate models, both predicted moderators positively correlated with HIV. HIV rates were higher in counties with more frequent references to risky activities (*B* = 0.21/0.28, *SE* = .04/.08, 95% CI = [0.14, 0.28]/ [0.12, 0.43], *z* = 3.50/5.97, both *p*s < .001), and higher STI prevalence (*B* = 0.31/0.45, *SE* = .06/.16, 95% CI = [0.19, 0.43]/[0.12, 0.77], *z* = 4.95/2.71, *p* < .001/.007).

Random state-level effects revealed that the effect of future tense was negative across all but two states (Illinois and Michigan) and was strong in most. Regression coefficients were close to zero (< |.05|) in only six states (Idaho, Iowa, Nebraska, Ohio, West Virginia, Wyoming). See Table 2.

Future Orientation Moderates Risk. In two sets of moderator models, future tense significantly interacted with risky language (*z* = -44.60) and STI prevalence (*z* = -28.25), both *p*s < .001. Decomposing the interaction by high ($z > 1$), moderate ($-1 < z \leq 1$) and low ($z \leq -1$)

future orientation, both risky language and STI prevalence had positive associations with HIV prevalence in counties with low rates of future tense (Risky language: *B* = 1.65, *SE* = .83, 95% CI = [0.03, 3.28], *z* = 2.00, *p* = .046; STI prevalence: *B* = 0.98, *SE* = .20, 95% CI = [0.59, 1.38], *z* = 4.89, *p* < .001). Only STI rates positively correlated with HIV rates in counties with moderate levels of future tense usage (*B* = 0.21, *SE* = .08, 95% CI = [0.06, 0.36], z = 2.81, *p* = .005), although the effect of was weaker than in counties with low future tense usage. Both risky language and STI rates were uncorrelated with HIV rates, both *p*s > .19, in counties with high rates of future tense. Results suggest that future-oriented language may buffer communities from both sexual and behavioral risk factors for HIV transmission. (See Figures 1a and 1b.)

Note that the association between county-level STI rates and HIV is reduced to nonsignificance for future-oriented counties only when we control for the Black population, foreign-born population, and population density. As a single predictor (with no controls) in the same multilevel model, STI rates significantly positively correlate with HIV rates even in high future-orientation counties, *B* = .21, *SE* = .10, CI = [.01, .41], *p* = .036. Risky language, on the other hand, is unrelated to HIV rates at the highest level of future orientation even in a single-predictor model. Thus, it is not the case that sex risk is no longer associated with HIV in highly future oriented counties. Rather, the association between STI and HIV rates is substantially weakened in those counties.

Differential Language Analysis

DLA produced two word clouds depicting the words that most strongly correlated with HIV prevalence (Figures 2a and 2b). Both analyses controlled for percentage of Black population and population density, the two best predictors of HIV prevalence. We analyzed only the 400 most populous counties for this set of analyses in order to focus on the highest-risk counties and further avoid conflating urban versus rural language differences with HIV prevalence. In the word clouds, larger words reflect stronger correlations, whereas darker words indicate greater frequency.

The negative cloud is the most revealing of the two. Several of its words refer not only to the future tense (e.g., *would be*), but to preparation for the future (e.g., *tomorrow, bed*) and thinking about alternate possible futures (e.g., *hopefully*, *might*). For the positive cloud, many of the words appeared to be references to urban nightlife (e.g., *lounge*, *dj*) and consumerism (e.g., *fashion*, *ads*) or are slang (e.g., *yoo*, *sa*). Both topics are problematic for individuals higher in impulsiveness and delay discounting, offering momentary temptations that result in increased health risk or decreased financial security in the future.

Additional Analyses

There are a huge number of structural and demographic variables for which county-level data are available. Although we do not have the space here to explore every possible moderator of the association between future orientation and HIV prevalence, two of the broadest potential moderators are explored below.

Location x future orientation. Beyond estimating state-level effects (Table 2), we also explored how longitude and latitude relate to the association between future orientation and HIV prevalence. Single pairs of coordinates for each county were taken from the United States Census Bureau's (2014) Gazetteer, which provides interpolated centroids of county border polygons. In order to model the linear effect of future orientation on HIV rates as a function of geography, we regressed HIV rates onto smoothing splines for the effects of longitude, latitude, and future orientation in a Poisson general additive model (GAM) regression. Control variables were the same as above (Black and foreign born population percentages, and population density).

Because the purpose of this analysis was to test how geographic location relates to the effect of future orientation on HIV rates, we did not nest within states in a multilevel model.

The smooth term for longitude, latitude, and future was significant ($X^2 = 33{,}269$, $p < .001$), and each of the control variables remained significant as well (all $z > 50$, all $p < .001$). Visualizing the smoothing splines (Figures 3a and 3b) illustrates that, consistent with the state-level effects displayed in Table 2, the linear effect of future orientation on HIV rates was strongest in counties with middle and lower (southern) latitudes, and weakest in counties that had both low (eastern) longitudes and high rates of future tense.

Population density x future orientation. Twitter users are slightly younger and more urban than the average American (Duggan & Brenner, 2013); therefore, it may be the case that our results primarily apply to more urban counties. To test whether population density moderated the effect of future orientation on HIV, we regressed HIV rates on future orientation, the three structural control variables used earlier (population density and percentage Black and foreign born population), and the future x population density interaction in a multilevel Poisson regression model that included random intercepts for states. The future x population density interaction was significant, $p < .001$, and was driven by the fact that the effect of future orientation is weaker in counties with population densities at or below the median ($B = -0.05$, $SE = .01$, $z = -9.76$, $p < .001$) than in counties with higher population densities ($B = -0.26$, $SE = .004$, $z = -62.14$, $p < .001$). However, the effect of future-oriented language remained significant at each level of population density.

Discussion

Future oriented messages on social media may reflect social norms to plan for, value, and generally think about the future. Thus, we set out to study whether these messages may correlate with HIV infections. Using out-of-sample prediction, we found that future orientation on Twitter negatively correlates with counties' HIV prevalence after controlling for traditionally strong predictors of HIV rates, including population density and ethnic composition. Furthermore, results of two sets of moderator analyses were consistent with the view of future orientation as a buffer against HIV risk, with STI rates and tweeted references to risky activities showing weaker correlations with HIV rates in counties that had higher rates of future-oriented messages.

Notably, a person would not need to use future-oriented language or think about the future to benefit from living in a community where thinking about the future is common or normative. Both protective and risky behavior patterns tend to be contagious (Christakis & Fowler, 2013), and individuals often conform to social norms or pursue socially valued goals without conscious deliberation (Custers & Aarts, 2010). Thus, residents of healthier communities may engage in protective health behaviors without internalizing the motivations that can lead individuals to independently engage in those behaviors. Following the lead of previous peer-education interventions that have trained participants to deliver intervention contents (e.g., how to bleach needles) to the general community after the intervention ended (Tobin, Kuramoto, Davey-Rothwell, & Latkin, 2011), future interventions may leverage the power of social contagion by training influential individuals in at-risk communities to both think about the future and discuss plans for the future with others on and off of social media sites.

Based on these data alone, we cannot know what kinds of socially transmissible health behaviors might accompany future orientation. Prior research regarding thinking about the future and delay discounting suggests that future orientation will increase impulse control and preventive health behaviors, such as exercising and wearing seat belts (see Daugherty & Brase, 2010). If it is true that future orientation manifests in specific health behaviors that help to reduce rates of HIV transmission in communities, such as regularly getting tested for HIV or using condoms, research supports the prediction that those behaviors would spread through social networks without requiring those who are being influenced to themselves adopt a more future-oriented thinking style. Future research should address whether people who mimic others' future-focused behaviors themselves begin to think more about the future as a result, and whether this kind of social contagion can be observed in communities over time.

Limitations and Future Research

Although consistent with the existing experimental research on future time perspective and related processes, such as delay discounting (e.g., Daugherty & Brase, 2010), our findings are correlational and thus do not allow us to assess whether future orientation itself has a causal influence on health behaviors or HIV rates. Both predictors and outcomes in our models were also assessed in the same time period. To work around the fact that naturalistic language use may be impossible to manipulate at the level of entire communities, future studies may track future-oriented language in counties over time to determine whether naturally occurring increases in talking about the future predict later increases in protective behaviors, such as using condoms, getting tested for HIV, or engaging in other more broadly healthy behaviors. These longitudinal analyses would bring us closer to our goal of tracking future-oriented language, along with moderators such as risky language, in order to forecast the locations of new HIV outbreaks and intervene before they spread throughout the community.

Using naturalistic Twitter data also prevents us from testing whether future orientation itself is beneficial, or whether merely living in a community of people who tend to think about and plan for the future offers health benefits. At the structural level alone, future oriented communities may be more likely to have infrastructure in place that promotes the health of its residents. For example, regardless of their own time perspective, presumably most people would benefit from living in a place where future-oriented community leaders have increased the accessibility of inexpensive preventive healthcare, free condoms (or, for injection drug users, needles), and information about preventing HIV and other diseases. Future studies would benefit from experimentally testing whether it is individuals' own future orientation or the time perspective of those surrounding them that affects protective and risky health behaviors, at the level of both specific social interactions as well as communities.

Another potential limitation is the fact that language used on Twitter deviates from everyday communicative language use in several ways. Computer-mediated communication such as Twitter or email differs from face-to-face communication in that it is asynchronous and editable (Hancock, 2007). Twitter further limits the number of characters that people can use and, depending on how many people are following a person's Twitter account, tweets are often very public. Each of these factors increases the likelihood that the language we analyzed was heavily edited or otherwise biased by individuals' desires to please their followers. Indeed, analyses of public versus private writing suggest that people writing publicly tend to screen socially undesirable content such as swearing or negative emotion words (Mehl, Robbins, & Holleran, 2012; Robbins et al., 2011), an effect that may have blunted our risky language dictionary's ability to assess the degree to which people in a community engage in risky behavior. However, importantly, LIWC's future dictionary is made up entirely of function words such as *will* and *should*. Function words are difficult to self-regulate and, because they have very little meaning outside the context of a sentence, are unlikely targets of editing (Bell, Brenier, Gregory, Girand, & Jurafsky, 2009). Thus, we are confident that the future category, at least, is relatively immune to biases introduced by the Twitter medium and offers an accurate assessment of the degree to which a community tends to focus on the future.

Finally, the Twitter corpus that we analyzed dates from late 2009 and early 2010. Twitter was popular at that time but less ubiquitous than it is now. The Twitter users in our sample were therefore relatively early adopters who were likely more socially active and educated than the average US resident (Sheldon, 2012). Demographic data for that period also suggest that the predominance of 18-29-year-olds on the site was more pronounced then than it is now (Lenhart, Purcell, Smith, & Zickuhr, 2010). Although more educated people tend to be less impulsive and better at planning for the future, younger and more extraverted individuals show the opposite pattern (Hirsh, Morisano, & Peterson, 2008; Reimers, Maylor, Stewart, & Chater, 2009). These tendencies, coupled with the fact that individuals in early adulthood are at the greatest risk of HIV infection (CDC, 2014), suggest that early adopters of social media technology may be better bellwethers of communities' HIV risk than other segments of the population. Future research, perhaps using age and education information extracted from users' social media profiles, is needed to determine whether the predictive utility of future-oriented language varies across age or socioeconomic status.

Conclusion

Rates of HIV and other diseases that reflect social disparities are symptoms of risky communities. Some US counties, despite their high population density (and commensurately easy access to risky activities like drinking), have relatively low rates of HIV. We have shown that these counties' social media language reflects future-oriented thinking. In addition to reflecting lower impulsiveness and better health at the individual level, as past research has shown, a future-oriented time perspective appears to buffer HIV risk at the community level as well. Although these data alone cannot tell us whether future oriented thinking styles increase communities’ resilience or decrease rates of HIV transmission, the richness and breadth of the data at hand give us confidence that these findings have real-world relevance and may help forecast HIV outbreaks in the future.

References

AIDSVu (www.aidsvu.org). Emory University, Rollins School of Public Health. Accessed July 1, 2013.

Albarracín, D., Gillette, J. C., Earl, A. N., Glasman, L. R., Durantini, M. R., & Ho, M. H. (2005). A test of major assumptions about behavior change: A comprehensive look at the effects of passive and active HIV-prevention interventions since the beginning of the epidemic. *Psychological Bulletin, 131*, 856.

Back, M. D., Küfner, A. C., & Egloff, B. (2011). "Automatic or the people?" Anger on September 11, 2001, and lessons learned for the analysis of large digital data sets. *Psychological Science*, *22*, 837-838.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*, 92–111. doi:10.1016/j.jml.2008.06.003

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

Boyd, J. N., & Zimbardo, P. G. (2005). Time perspective, health, and risk taking. In A. Strathman & J. Joireman (Eds.), *Understanding behavior in the context of time: Theory, research, and application* (pp. 85-107). Mahwah, NJ, US: Lawrence Erlbaum.

Center for Disease Control (CDC). (2014). Estimated HIV incidence in the United States, 2007–2010. *HIV Surveillance Supplemental Report 2012, 17*.

Chen, M. K. (2013). The effect of language on economic behavior: Evidence from savings rates, health behaviors, and retirement assets. *The American Economic Review*, *103*, 690-731.

Christakis, N. A., & Fowler, J. H. (2013). Social contagion theory: Examining dynamic social networks and human behavior. *Statistics in Medicine*, *32*, 556-577.

Custers, R., & Aarts, H. (2010). The unconscious will: How the pursuit of goals operates outside of conscious awareness. *Science*, *329*, 47-50.

Daugherty, J. R., & Brase, G. L. (2010). Taking time to be healthy: Predicting health behaviors with delay discounting and time perspective. *Personality and Individual Differences, 48*, 202-207.

Duggan, M. & Brenner, J. (2013). A demographic portrait of users of various social media services - 2012. Pew Research Center, Washington, DC. Retrieved July 1, 2013, from http://www.pewinternet.org/files/old-media//Files/Reports/2013/PIP_Social-MediaUsers.pdf.

Ebreo, A., & Vining, J. (2001). How similar are recycling and waste reduction? Future orientation and reasons for reducing waste as predictors of self-reported behavior. *Environment and Behavior*, *33*, 424-448.

Fleming, D. T., & Wasserheit, J. N. (1999). From epidemiological synergy to public health policy and practice: The contribution of other sexually transmitted diseases to sexual transmission of HIV infection. *Sexually Transmitted Infections*, *75*, 3-17.

Green, L., Fry, A. F., & Myerson, J. (1994). Discounting of delayed rewards: A life-span comparison. *Psychological Science*, *5*, 33-36.

Hancock, J. T. (2007). Digital deception. In Joinson (Ed.), *Oxford handbook of internet psychology* (pp. 289-301). Oxford, UK: Oxford University Press.

Hecht, B., Hong, L., Suh, B., & Chi, E. H. (2011). Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 237-246). ACM.

Hirsh, J. B., Morisano, D., & Peterson, J. B. (2008). Delay discounting: Interactions between personality and cognitive ability. *Journal of Research in Personality*, *42*, 1646-1650.

Hoyle, R. H., & Sherrill, M. R. (2006). Future orientation in the self-system: Possible selves, self-regulation, and behavior. *Journal of Personality*, *74*, 1673-1696.

Lampos, V., & Cristianini, N. (2010). Tracking the flu pandemic by monitoring the social web. In *Cognitive Information Processing (CIP), 2010 2nd International Workshop on* (pp. 411-416). IEEE.

Lenhart, A., Purcell, K., Smith, A., & Zickuhr, K. (2010). Social media and mobile internet use among teens and young adults: Millennials. *Pew Internet & American Life Project*.

MacKillop, J., Amlung, M. T., Few, L. R., Ray, L. A., Sweet, L. H., & Munafò, M. R. (2011). Delayed reward discounting and addictive behavior: A meta-analysis. *Psychopharmacology, 216*, 305-321.

Markus H. & Nurius P. (1986). Possible selves. *American Psychologist, 41*, 954–69.

Mehl, M. R., Robbins, M. L., & Holleran, S. E. (2012). How taking a word for a word can be problematic: Context-dependent linguistic markers of extraversion and neuroticism. *Journal of Methods and Measurement in the Social Sciences*, *3*, 30-50.

Mischel, H. N., & Mischel, W. (1983). The development of children's knowledge of self-control strategies. *Child Development*, *54*, 603-619.

Nurmi, J. E. (1987). Age, sex, social class, and quality of family interaction as determinants of adolescents' future orientation: A developmental task interpretation. *Adolescence, 22,* 977-991.

Oyserman, D., Bybee, D., & Terry, K. (2006). Possible selves and academic outcomes: How and when possible selves impel action. *Journal of Personality and Social Psychology*, *91*, 188–204. doi:10.1037/0022-3514.91.1.188

Pennebaker, J.W., Booth, R.E., & Francis, M.E. (2007). Linguistic Inquiry and Word Count: LIWC2007. Austin, TX: LIWC.net. [computer software]

Picard, R. R., & Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association*, *79*, 575-583.

Preis, T., Moat, H. S., Stanley, H. E., & Bishop, S. R. (2012). Quantifying the advantage of looking forward. *Scientific Reports*, *2*.

Prejean, J., Song, R., Hernandez, A., Ziebell, R., Green, T., Walker, F., Lin, L. S., An, Q., Mermin, J. Lansky, A., Hall, H. I., & HIV Incidence Surveillance Group. (2011). Estimated HIV incidence in the United States, 2006–2009. *PloS one*, *6*, e17502.

Random.org. Retrieved April 15, 2014, from http://www.random.org/sequences/.

Reimers, S., Maylor, E. A., Stewart, N., & Chater, N. (2009). Associations between a one-shot delay discounting measure and age, income, education and real-world impulsive behavior. *Personality and Individual Differences*, *47*, 973-978.

Robbins, M. L., Focella, E. S., Kasle, S., López, A. M., Weihs, K. L., & Mehl, M. R. (2011). Naturalistically observed swearing, emotional support, and depressive symptoms in women coping with illness. *Health Psychology*, *30*, 789.

Rothspan, S., & Read, S. J. (1996). Present versus future time perspective and HIV risk among heterosexual college students. *Health Psychology, 15,* 131-134.

Sadilek, A., Kautz, H. A., & Silenzio, V. (2012, July). Predicting disease transmission from geo-tagged micro-blog data. In *AAAI.*

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Agrawal, M., Park, G. J., Lakshmikanth, S. K., Jha, S., Seligman, M. E. P., Ungar, L., & Lucas, R. E. (2013). Characterizing geographic variation in well-being using Tweets. *In Seventh International AAAI Conference on Weblogs and Social Media (ICWSM 2013)*: Boston, MA.

Sheldon, P. (2012). Profiling the non-users: Examination of life-position indicators, sensation seeking, shyness, and loneliness among users and non-users of social network sites. *Computers in Human Behavior*, *28*, 1960-1965.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*, 24-54.

Tobin, K. E., Kuramoto, S. J., Davey-Rothwell, M. A., & Latkin, C. A. (2011). The STEP into Action study: A peer-based, personal risk network-focused HIV prevention intervention with injection drug users in Baltimore, Maryland. *Addiction*, *106*, 366-375.

United States Census Bureau. (2014). National Counties Gazetteer File [Data file]. Retrieved from https://www.census.gov/geo/maps-data/data/gazetteer2014.html.

Weller, R. E., Cook III, E. W., Avsar, K. B., & Cox, J. E. (2008). Obese women show greater delay discounting than healthy-weight women. *Appetite*, *51*, 563-569.

Whitaker, K. L., Good, A., Miles, A., Robb, K., Wardle, J., & von Wagner, C. (2011). Socioeconomic inequalities in colorectal cancer screening uptake: Does time perspective play a role? *Health Psychology*, *30*, 702.

Young, S. D., Rivers, C., & Lewis, B. (2014). Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. *Preventive Medicine, 63,* 112-115. doi: http://dx.doi.org/10.1016/j.ypmed.2014.01.024

Zimbardo, P. G., & Boyd, J. N. (1999). Putting time in perspective: A valid, reliable individual-differences metric. *Journal of Personality and Social Psychology*, *77*, 1271-1288. doi:10.1037/0022-3514.77.6.1271

Tables

Table 1

Dictionary Categories and Examples

Word category	Examples
Future	*could, gonna, may, might, must, ought, will, that'll, won't, shall*
Risky leisure	*blackjack, bong, casino, cigarettes, kegger, meth, skoal, shrooms, stoned, vodka*
Safe leisure	*beach, celebrate, chillin, jazz, karaoke, mall, museum, play, runner, scrapbook*

Note. "Future" is the future-tense verb category from LIWC2007 (Pennebaker, Booth, & Francis, 2007). The leisure categories are based on the leisure category from LIWC, which was expanded and subdivided into safe and risky leisure categories by trained raters for the present study.

Table 2

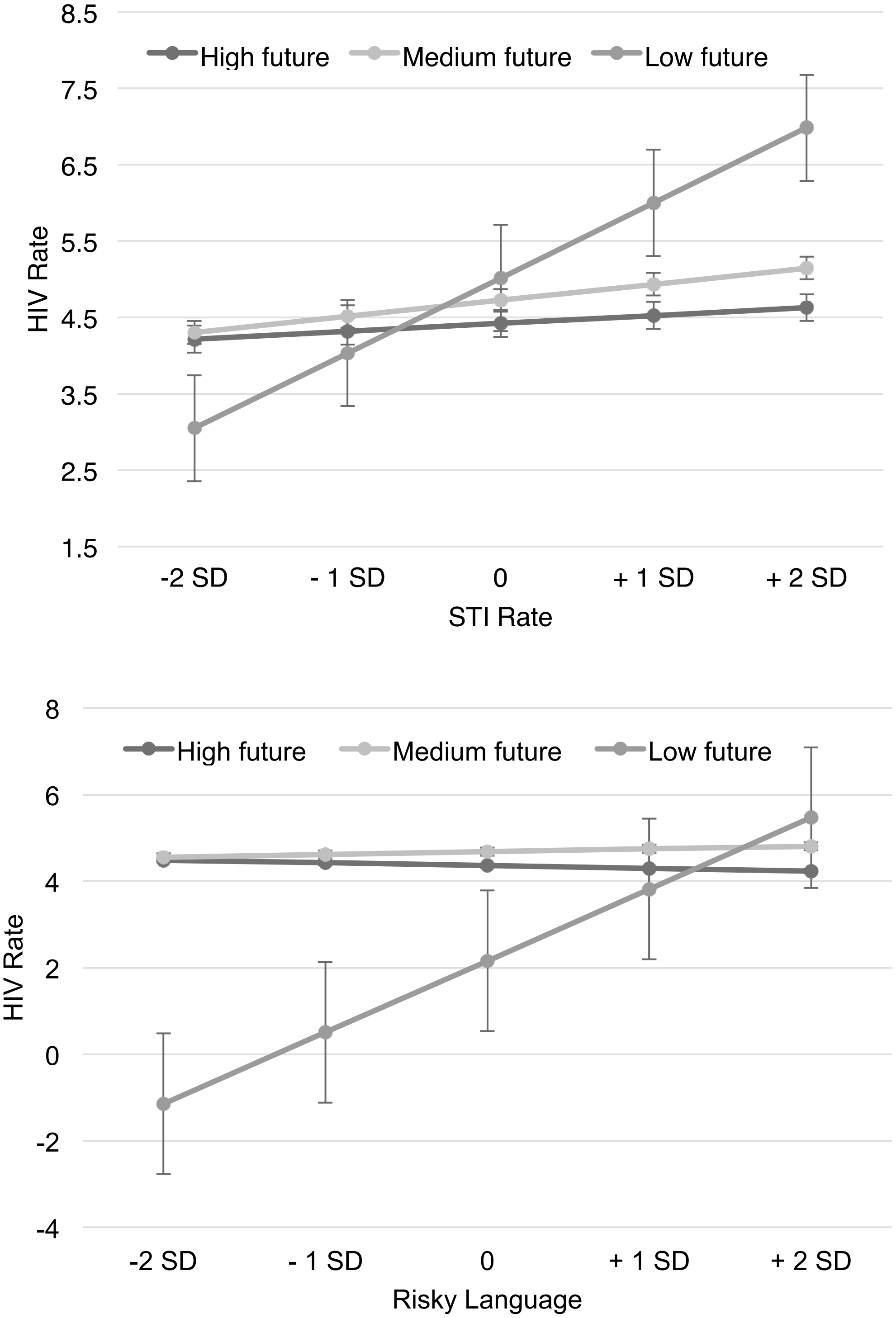
Random State-Level Effects of Future Orientation on HIV Prevalence

State	Intercept	*β*
Alabama	-0.13	-0.22
Arizona	0.04	-0.09
Arkansas	0.00	-0.12
California	-0.07	-0.26
Colorado	-0.03	-0.52
Connecticut	-0.06	-0.50
Delaware	-0.07	-0.49
Florida	-0.06	-0.85
Georgia	-0.02	-0.11
Hawaii	0.03	-0.14
Idaho	0.05	-0.02
Illinois	0.05	0.11
Indiana	0.06	-0.07
Iowa	0.04	0.02
Maryland	-0.66	-2.68
Massachusetts	0.04	-0.27
Michigan	0.02	0.06
Minnesota	0.02	-0.06
Mississippi	-0.14	-0.38
Missouri	0.10	-0.14
Montana	0.05	-0.05
Nebraska	0.05	0.00
Nevada	-0.02	-0.34
New Hampshire	0.02	-0.14
New Jersey	-0.11	-0.67
New Mexico	0.07	-0.07
New York	0.18	0.05
North Carolina	0.03	-0.21
Ohio	0.05	0.01
Oklahoma	0.04	-0.05
Oregon	-0.02	-0.24
Pennsylvania	0.08	-0.19
Rhode Island	-0.01	-0.16
South Carolina	0.01	-0.27
Tennessee	0.13	-0.08
Texas	-0.06	-0.12
Utah	0.03	-0.02
Vermont	0.05	-0.08
Virginia	0.07	-0.11
Washington	-0.03	-0.19

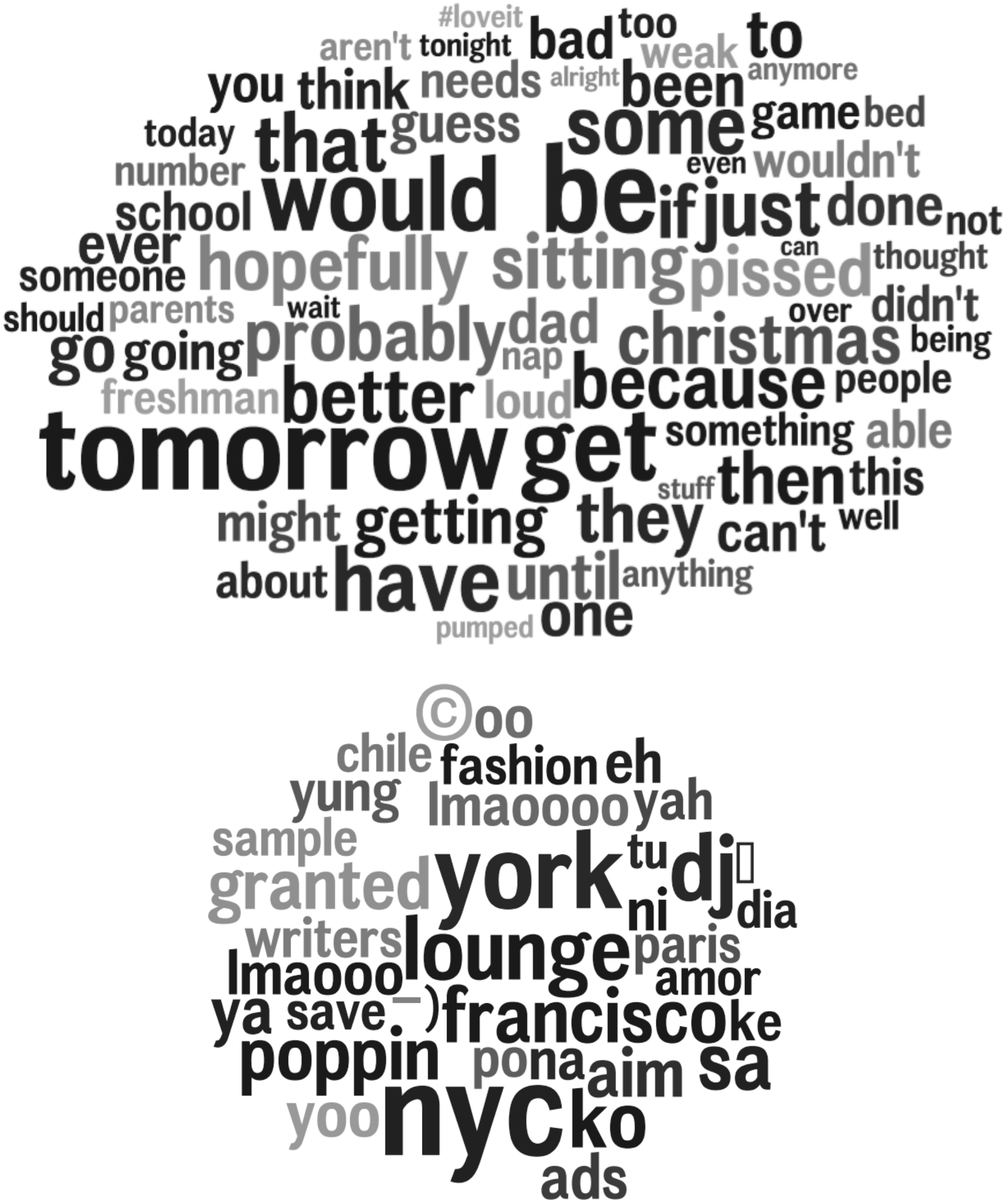
West Virginia	0.09	-0.01
Wisconsin	0.02	0.01
Wyoming	0.06	-0.03

Note. Random effects are from a multi-level model where slopes and intercepts for future oriented language were allowed to vary randomly between states.

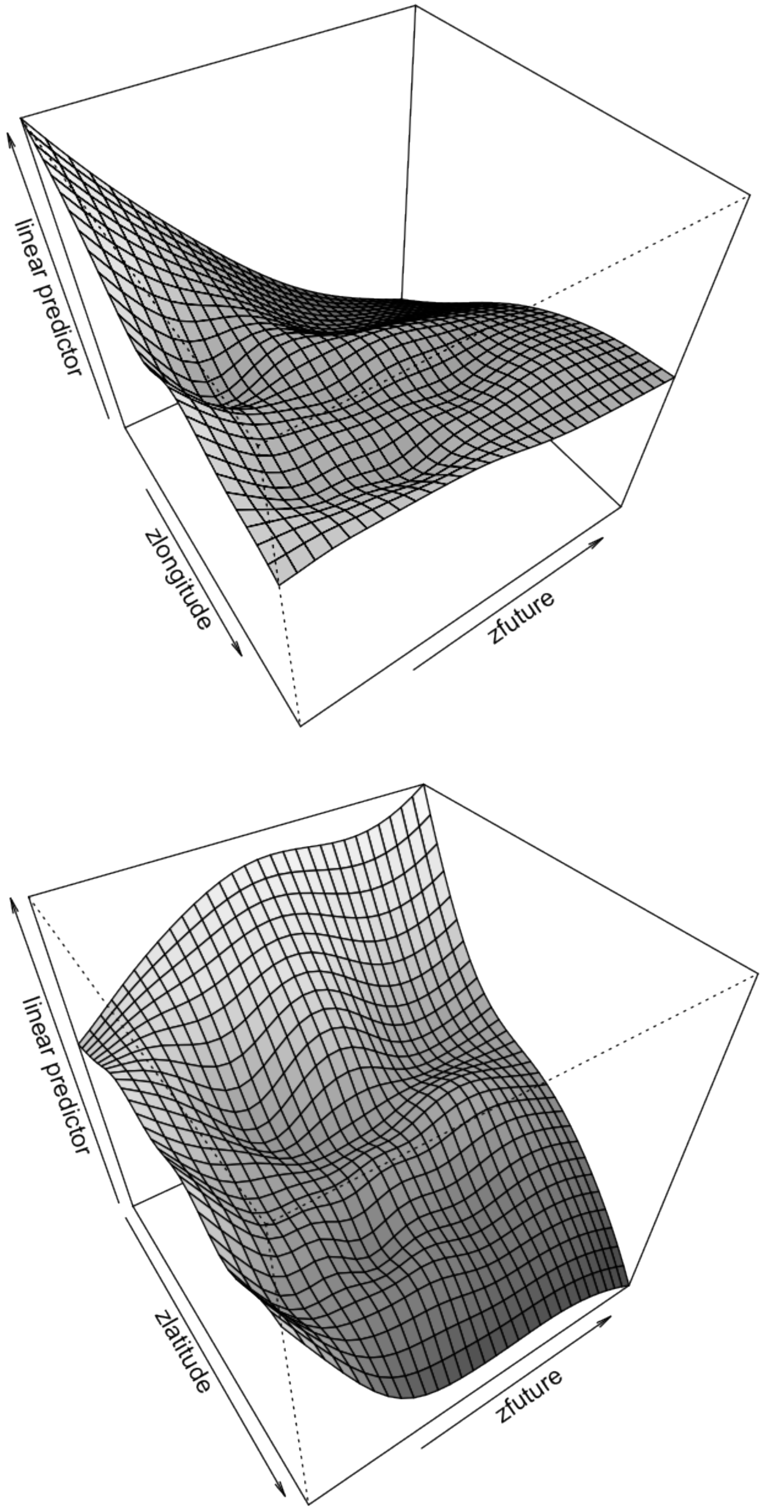
Figures



Figures 1a and 1b. Association between STI rate and HIV prevalence (top) and risky language and HIV prevalence (bottom) at high, moderate, and low levels of future tense usage.



Figures 2a and 2b. Words negatively (top) and positively (bottom) correlated with HIV prevalence in the 400 most populous US counties, controlling for percentage Black population and population density. For slang and nonstandard English words, *sa* = abbreviation for *esé*, an informal Latino form of address; *ko* = knocked out (or Filipino for *my*); *yoo* = extended version of the greeting *yo*; po = slang for vagina or short form of *poor*; *ke* = slang misspelling of the Spanish word *que* (or Indonesian for *to*).



Figures 3a and 3b. Smoothing splines modeling the linear effect of future orientation on HIV rates as a function of longitude (top), latitude (bottom), and percentage of future-oriented tweets. Darker = smaller effect sizes, higher longitudes = farther west, higher latitudes = farther north.

Footnote

[1] We use the term “natural language” in its broadest sense, meaning language that is produced by humans and used primarily for interpersonal communication or self-expression (in contrast with, for example, computer programming languages or other artificial languages).

[2] Conclusions were identical when controlling for socioeconomic status, as indicated by median income. When median income was included with future orientation and the other structural control variables as predictors of HIV prevalence in the multilevel Poisson regression model described earlier, median income failed to significantly predict HIV ($p = .121$) while future orientation remained a significant predictor ($z = -7.43$, $p < .001$). Thus, median income probably does not account for the negative association between future orientation and decreased HIV risk.