# Mining Twitter as a First Step toward Assessing the Adequacy of Gender Identification Terms on Intake Forms

**Amanda Hicks, PhD[1], William Hogan, MD, MS[2], Michael Rutherford, MS[1], Bradley Malin, PhD[3], Mengjun Xie, PhD[4], Christiane Fellbaum, PhD[5], Zhijun Yin, BE[3], Daniel Fabbri, PhD[3], Josh Hanna, MS[2], Jiang Bian, PhD, MS[2]**

**[1]University of Arkansas for Medical Sciences, Little Rock, AR; [2]University of Florida, Gainesville, FL; [3]Vanderbilt University, Nashville, TN; [4]University of Arkansas at Little Rock, Little Rock, AR; [5]Princeton University, Princeton, NJ**

**Abstract**

*The Institute of Medicine (IOM) recommends that health care providers collect data on gender identity. If these data are to be useful, they should utilize terms that characterize gender identity in a manner that is 1) sensitive to transgender and gender non-binary individuals (trans* people) and 2) semantically structured to render associated data meaningful to the health care professionals. We developed a set of tools and approaches for analyzing Twitter data as a basis for generating hypotheses on language used to identify gender and discuss gender-related issues across regions and population groups. We offer sample hypotheses regarding regional variations in the usage of certain terms such as 'genderqueer', 'genderfluid', and 'neutrois' and their usefulness as terms on intake forms. While these hypotheses cannot be directly validated with Twitter data alone, our data and tools help to formulate testable hypotheses and design future studies regarding the adequacy of gender identification terms on intake forms.*

## Introduction

The LGBT community is subject to a variety of health disparities. This is a result of a lack of meaningful data on LGBT populations as well as a lack of training and resources for clinicians to provide culturally competent care. Recent Institute of Medicine (IOM) recommendations to address these health disparities include (1) gathering data on sexual orientation and gender identity in Electronic Health Records (EHR) as part of the meaningful use objectives in EHRs, (2) developing standardization of sexual orientation and gender identity measures to facilitate synthesizing scientific knowledge about the health of sexual and gender minorities, and (3) supporting research to develop innovative methods of conducting research with small populations and to determine the best ways to collect information on LGBT minorities.[1]

While the IOM notes that data collection would be aided by standardized measures for sexual orientation and gender identity, their report also emphasizes that defining sexual orientation and gender nonconformity is a challenge since these are multifaceted concepts. The use of terminology that is familiar to the participant has been shown to improve response rates.[1,2] However, based on the limited research available, there is some evidence[3,17,18] to suggest that consumer vocabulary for self-identifying gender and sexual orientation varies by community. There is clear evidence of lexical variation associated with geography in linguistics studies.[19,20,21] Also, through discussions with members of the trans* community and health care providers at LGBT clinics across the country, we have learned that new terms are frequently being coined to describe gender identity and that the connotations of existing terms may vary by community.

There is documentation of variation of terms used to describe sexual orientation across communities.[25] There is also variation in the meaning of terms between individuals who consider themselves part of the sexual minority (e.g., lesbian, gay, or bisexual) and those who do not (e.g., straight or heterosexual).[26] For example, self-identifying members of a sexual minority use 'lesbian' to refer to women who are *primarily* attracted to other women, but others tend to use 'lesbian' more broadly to refer to a woman who has experienced *any* sexual attraction or sexual activity with another woman.[26] This raises the question of whether there is similar variation in the meanings of terms used to describe transgender identity. However, data addressing variations of gender identity terms and their meanings is lacking. This is significant for the development of good intake forms; if there is significant lexico-semantic variation of gender identity terms, then a single, universal standard intake form may result in a lower response rate than intake forms that are community specific.

A number of organizations have attempted to address the question of how to ask patients about their gender identity. A summary of these approaches can be found in the GenIUSS Report by the Williams Institute.[4] The most promising is a two-step format recommended by the UCSF Center of Excellence for Transgender Health. This approach first asks patients about their gender identity and then their sex assigned at birth.[4] However, this research addresses the form of the question, not the specific items used to present gender-identity options that ought to be available on the form. The language used in the gender identity question varies across forms from different healthcare organizations. Table 1 contains the choices from the sample forms of three institutions: 1) Fenway Health, 2) UCSF, and 3) the Williams Institute. Although the two-step format has been field tested in Michigan by the Fenway Institute,[16] it is not clear to what extent the terms on these forms represent the identity terms used by transgendered, non-binary, and/or gender-variant people (trans*[i]) across the United States. The result is that there are still outstanding questions regarding which terms are optimal for intake forms and whether a single, universal standard terminology will suffice for all trans* communities.

| **Fenway Health Intake Form** | **UCSF Center of Excellence Sample form** | **GenIUSS Sample Form** |
|---|---|---|
| Male<br>Female<br>Genderqeer or not exclusively male or female | Male<br>Female<br>Transgender Male/Transman/FTM<br>Transgender Female/Transwoman/MTF<br>Genderqueer<br>Additional category (please specify)<br>Decline to Answer | Male<br>Female<br>Trans Male/Trans Man<br>Trans Female/ Trans Woman<br>Genderqueer<br>Different Identity (please state) |

**Table 1.** Gender identity terms found on various intake forms.

User generated content on social media, such as Twitter, is a valuable resource because it can provide a source for gleaning information about people's daily life to answer scientific questions. We believe this source can produce a data set that can contribute to the IOM priority area to study social influences on LGBT health and to the IOM recommendation to develop innovative methods for conducting research on small populations.[1] Mining social networking resources produces data sets that can be used to investigate social influences of health concerns among transgender persons.

Our goal is to build a data set and visualization tools that can be used as a basis to generate hypotheses for further testing to guide the development of gender identity questions on intake forms. Our process for building these tools was as follows. We first examined which terms are currently used to describe transgender identity on Twitter. Based on existing research on linguistic variation in social media,[22] we hypothesize that the usage of gender identification terms varies by geographical region. Then we geotagged the tweets by US state, classified tweets as authored by self-identifying transgender users, and created a co-occurrence network and term frequency counts to support hypothesis generation with data visualization tools. These co-occurrence counts and frequency counts will form the basis of distributional similarity metrics in future research to help determine a) whether different terms are synonyms; and b) whether some terms are polysemous.[27]

Our approach is consistent with the intersectional perspective recommended by the IOM. The intersectional approach considers sub-populations of the LGBT community based on several orthogonal factors, such as ethnicity and geographical region. Furthermore, the resulting data set can be used to address demographic research, social influences on health, and transgender specific health needs — three of the five priority research areas.[1]

While the results from Twitter mining do not always yield language that is appropriate in the context of clinical care and research — for example, there is a significant quantity of advertisements for sex work on Twitter and discussions of gender-related slurs — Twitter has the potential to provide a comprehensive snap-shot of the language used by self-identified trans* individuals.

[i] For brevity we refer to transgendered, gender non-binary and gender-variant people by the term trans*.

Another goal of this paper is to establish a set of best practices for dealing with social media for extracting useful biomedical knowledge, which can help produce data on small populations through unfettered access to such a "Big Data" source (over 500 million tweets per day[ii]).

**Methods**

The general idea underlying our approach is to identify tweets that are relevant to the discussion of trans* related issues, and then examine the variations in language used for gender identification by different communities, that is, by population (trans* people vs. the general public) and by geographical location (U.S. states). The analysis workflow consists of five main steps, as depicted in Figure 1: 1) collect tweets that are potentially related to discussions about gender identification; 2) preprocess and geotag tweets with their corresponding U.S. state; 3) build supervised classification models based on textual features in the tweets to a) filter out irrelevant tweets[iii] and b) find people who are self-identified as trans*; 4) collect relevant (both self-identifying trans* and people in the general public who discussed trans* related issues) users' Twitter timelines[iv]; and 5) compare the usage of gender identification terms by geographical locations (i.e., by U.S. states) and by population groups (trans* people vs. the general public). We also leverage a number of visualization techniques to provide straightforward and easy-to-understand visual representations – word clouds, co-occurrence matrices, and network graphs – to substantiate our findings. In the following sections, we describe each step and the basic procedures in further detail.

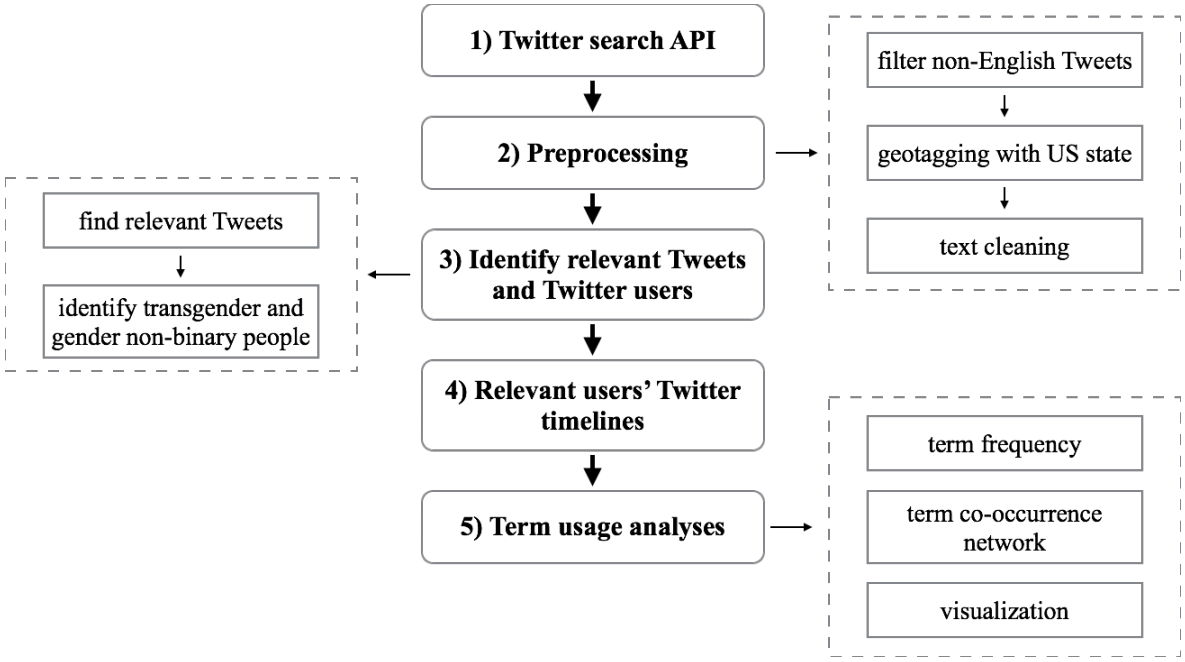



**Figure 1.** The general analysis workflow consists of five steps: 1) collect relevant tweets using the Twitter search API with a search term list; 2) preprocess the collected data to filter out non-English tweets and geotag based on user profiles; 3) build classification models to identify relevant tweets and Twitter users; 4) collect relevant users' Twitter timelines; and 5) analyze the usage of keyterms through comparing term frequencies and co-occurrences.

*a) Data collection through the Twitter search API*

We developed a set of Python scripts leveraging the *twython*[v] library for accessing the Twitter APIs. We designed our Python crawler, *tweetf0rm*[vi], to handle various potential runtime exceptions (e.g., the crawler will recover from a

[ii] http://www.internetlivestats.com/twitter-statistics/

[iii] Some of the search terms are ambiguous and their meanings are context dependent. For example, the tweet "That Hot Pocket is full of trans fats" is not related to discussions of gender identification even through it contains the keyword. To account, we built a binary classifier to determine how likely that a tweet is relevant to the discussion of gender identification and remove those that are likely to be irrelevant from the corpus, as described below.

[iv] The timeline of a Twitter user corresponds to all of their tweets in chronological order.

[v] https://github.com/ryanmcgrath/twython

system failure automatically and pause collection when it reaches the Twitter API rate limits[vii]) and distribute the workload across multiple Amazon EC2 instances. The data collection process began with a list of keywords (i.e., search terms) mainly related to gender identification such as 'transwomen', 'genderqueer', and 'transmasculine'. We have also included a number of other keywords that could indicate relevance of the tweets to trans* discussions such as 'testosterone' (a drug often used as part of the hormone replacement therapy for transgender individuals) and 'gender reassignment surgery'. To ensure coverage, we considered the base forms of these terms as well as their spelling variations, such as 'transwomen', 'trans-women', and 'trans women'.

Additionally, we found that a number of hashtags (i.e., patterns that start with '#' to mark topics in a tweet and often used by Twitter users to categorize the messages), such as '#iamnonbinary' and '#iamtrans', are good search terms with a low false positive rate for identifying tweets relevant to our study. To develop a list of search terms, we started with 'transgender' and 'trans' as seed terms which we used as search terms on Twitter and manually compiled a list of co-occurring terms that are in the domain of trans* gender-identification. We next iterated this process until we were no longer accumulating new terms. Then we manually examined the collected tweets to determine the quality of these terms as search terms. Through an iterative process, we removed terms where the majority of the returned tweets were false positive and added new relevant keywords that discovered in the collected tweets.

*b) Data preprocessing and geotagging*

We preprocessed the collected data to eliminate tweets that 1) were not written in English or 2) those for which we could not determine the geographical location of the user. For language detection, we leveraged the Twitter API metadata directly, which includes a 'lang' attribute specifying the language that the tweet was written in.[viii] For geotagging, we extracted the 'location' field, part of a user's profile, and attempted to assign a U.S. state to each tweet accordingly. Specifically, we searched each location field for a number of lexical patterns indicating the location of the user such as the name of a state (e.g., Arkansas or Florida), or a city name in combination with a state name or state abbreviation in various possible formats (e.g., "——, fl" or "——, florida" or "——, fl, usa"). Self-reported locations are often nonsensical[5] (e.g., "wonder land" or "up in the sky"), but strict patterns produced good matches and helped to reduce the number of false positives.

Notably Twitter also provides the ability to attach geocodes (i.e., latitude and longitude) to a user's profile and to each tweet. Yet, since geolocation needs to be enabled explicitly by the user as well as requires the user to have a device that is capable of capturing geocodes (e.g., a mobile phone with GPS turned on), very few tweets we have collected have this information. This is consistent with findings from previous studies.[6,7] If the 'location' field was missing in a user's profile, but the 'geo' attribute was available, we attempted to resolve the location of the user through reverse geocoding via the publicly available GeoNames geographical database.[ix] However, we did not use the geocodes attached to each individual tweet[x] since it is possible that a user was traveling away from their home state, in which case the geocodes attached to the tweets would be different from those on their profile. For our study, we geotagged the tweets based on where the user is from, not where the user is traveling temporarily. However, we do consider the scenario where a user permanently moved from one state to another reflected as a change in the 'location' field of a user's profile.

We have also made a number of other efforts to clean up the tweets including: 1) fixing Unicode text using *ftfy*[xi]; 2) removing mentions (i.e., indicating conversations in a tweet, starts with '@' followed by a username); and 3) eliminating hyperlinks. However, we did retain hashtags as they indicate topics and categories of the tweets and may contribute to the vocabulary of trans* related discussions.

*c) Classification models for finding relevant tweets and Twitter users*

Even though a tweet contains one or more of the search keywords, the tweet may not be relevant to our study due to the ambiguity of the search terms. The meanings of many search terms are context dependent. For example, the term

---

[vi] https://github.com/bianjiang/tweetf0rm
[vii] https://dev.twitter.com/rest/public/rate-limiting
[viii] https://blog.twitter.com/2013/introducing-new-metadata-for-tweets
[ix] http://www.geonames.org/
[x] In Twitter, geocoding can be either at user-level or at individual tweet-level.
[xi] https://github.com/LuminosoInsight/python-ftfy

'trans' could also mean "trans fat" or "transmission", depending on the context of the sentence. Since we are interested only in tweets where 'trans' means "transgender", we built a binary classifier to distinguish tweets that are relevant vs. irrelevant to the discussion of gender-related issues. Further, we want to examine whether there are any differences in the terminology used across trans* communities. Thus, we built a second binary classifier to discover people who are self-identified as trans*.

The mechanisms of both classifiers are essentially the same. We first converted each tweet into a feature vector using the term frequency-inverse document frequency (tf-idf) scheme[8] and then trained the classifiers using a random forest.[9] We manually annotated 6,058 tweets to obtain a training sample. All tweets were read by three people and each tweet was assigned one of three labels: 'irrelevant' (661 tweets), 'relevant but NOT self-identifying' (4,619 tweets), and 'relevant AND self-identifying' (778 tweets). When disagreements between the three annotators occurred, we used the majority rule to determine the final label. Although the three labels are mutually exclusive in the sense that only one label is assigned to each tweet, self-identifying tweets are inherently relevant tweets. Therefore, in building the disambiguation classifier, we treated relevant tweets (both self-identifying and not self-identifying) as positive samples and irrelevant tweets as negative samples. In building the second classifier to identify the trans* population, we treated self-identifying tweets as positive and the remainder of the relevant tweets as negative. We followed standard machine learning best practices (e.g., use 10-fold cross-validation to find the best model parameters—the number of trees in the forest for the random forest model, and for both classifiers the best parameters are 110) to ensure these classifiers are of high quality. The prediction accuracy for finding irrelevant tweets is 97.4% (precision: 0.970; recall: 0.766), and the accuracy for identifying trans* people is 87.8% (precision: 0.741; recall: 0.261).

*d) Collect relevant users' Twitter timelines*

Further, we expanded our corpus to include all the tweets posted by the users who were classified as trans*. The motivation for collecting relevant users' Twitter timelines is two-fold. First, the Twitter search API only returns recent tweets[xii], but a user could have posted discussions related to trans* issues beyond the search limit. Second, our list of search terms does not contain all of the keyterms of interest[xiii], such that a user could have posted discussions that contain one or more terms that are not search terms. A user's Twitter timeline can be collected using Twitter's 'statuses/user_timeline' API. However, the Twitter user timeline API only return up to 3,200 of a user's most recent tweets. Therefore, our crawling tool continuously monitors all relevant users' timelines to collect data beyond the 3,200 limit. Note that our approach cannot go beyond the limit for historical data, but rather is a way to circumvent the limit for future tweets.

*e) Generating term frequency and co-occurrence networks*

From the collected tweets, we calculated the term frequency statistics of the keyterms that we are interested in at both national and state level. Moreover, to provide a fair state-by-state comparison, the term frequency statistics were normalized by the number of total tweets of each state. Comparing the term frequency statistics can suggest regional differences in terminology, which in turn can lead to focused hypotheses for further investigations.

Furthermore, we produced co-occurrence networks of the keyterms hoping to discover semantic proximities and the latent structure among them.[10,11,12] We formalize a key term co-occurrence network as an undirected weighted graph, $G = (V, E)$, where each term is a vertex or node ($v_i$). If two terms co-occurred (in any order) in the same Twitter message, we drew an edge or link ($e_{ij}$) between the two term nodes ($v_i$ and $v_j$), such that the weight ($w_{ij}$) of the edge is set to the number of co-occurrences in all tweets posted by the users of interest. We constructed two co-occurrence networks for each state—one representing the trans* population and the other for the general public (including trans* people).

To assist in the presentation of the results, we built a number of web-based visualizations (http://bianjiang.github.io/twitter-language-on-transgender/). In particular, we used word clouds to depict the

[xii] Twitter does not release the information of their search algorithm to determine how many days of data will be returned before the day an inquiry is submitted, but our data set suggests it is around 14 days.

[xiii] Our search term list is rather restrictive, and does not contain all the gender identification terms that we are interested in to eliminate too many false positives. We removed a term from the search term list when the majority of the tweets it returned are irrelevant. For example, we found that "ftm" ("female-to-male", but could also mean "first time mom") performed extremely poor.

representative keyterms; and built interactive network visualizations using a physically-based force-directed graph layout with the Scalable Vector Graphics (SVG)—a language for building rich graphical content,[13] and d3—a JavaScript library for manipulating SVG objects.[14]

**Results**

We collected over 31 million tweets matching the search queries during a 49-day period from January 17, 2015 to March 6, 2015 inclusive. Out of the collected tweets, about 11 million tweets (36.1%) were in English. We were able to extract location information for 141,400 tweets (1.24% of English tweets from 57,997 unique users), which we retained for further processing. Next, we applied the two developed classifiers. We eliminated the tweets that were deemed irrelevant (5,685 tweets from 1,899 users). From the rest of the data set, 56,098 Twitter users were classified as relevant, of which 1,129 users were classified as self-identifying trans*. In addition to the data we collected using the search API, we crawled more than 154 million tweets from the 56,098 relevant Twitter users' timelines. Out of the 154 million tweets, 532,682 Twitter messages contain one or more the keyterms of our interest. These 500k tweets represent the corpus we used for language usage analysis.

Table 2 shows the top ten most frequently used keywords across the US on Twitter by trans* people vs. the general public. We present the data on the percentage scale to make the results comparable between the two population groups. In the table the term 'trans' occurs frequently because it is part of other keywords (e.g., 'trans people' and 'trans woman') that we are interested in. For the same reason, 'trans' co-occurred frequently with terms like 'trans people' and 'trans woman'. For the purpose of better presentation, we removed any top ranked co-occurrence pairs that contain the term 'trans' in Table 2.

As reported in Table 2, the most frequently used terms are similar between users classified as trans* and the general public on the national level. The Spearman's rank correlation coefficient of the term frequency lists (i.e., the general public vs. trans* people) yields a value of $0.943$ (with a two-sided p-value of $8.38 \times 10^{-47} < .01$ significance level) indicating the two lists are highly correlated. On the national level there is a common vocabulary invoked to discuss gender-related issues online.

| Rank | General Public Terms | | Trans* People Terms | |
|---|---|---|---|---|
| | Term Frequency | Co-occurring | Term Frequency | Co-occurring |
| **1** | trans (*32.05%) | #tgirl, shemale | trans (34.73%) | #tgirl, tranny |
| **2** | transgender (19.71%) | #tgirl, sissy | transgender (14.78%) | shemale, tranny |
| **3** | cis (6.81%) | shemale, sissy | cis (7.35%) | #tgirl, shemale |
| **4** | shemale (3.78%) | shemale, tranny | shemale (4.24%) | #tgirl, sissy |
| **5** | gender (3.51%) | gender, transgender | trans people (3.48%) | gender, transgender |
| **6** | transphobia (3.12%) | #tgirl, tranny | transphobia (3.19%) | ladyboy, shemale |
| **7** | tranny (3.11%) | ladyboy, shemale | tranny (2.88%) | shemale, sissy |
| **8** | trans people (2.78%) | #tgirl, ladyboy | gender (2.83%) | ladyboy, tranny |
| **9** | #tgirl (2.15%) | gender, gender binary | #tgirl (2.57%) | cis, gender |
| **10** | trans woman (1.96%) | ladybody, tranny | transsexual (2.30%) | dysphoria, gender |

**Table 2.** The top ten terms and co-occurring terms tweeted across the United States by the general public vs. trans* for gender identification and discussions of gender-related issues. (**The number in the parenthesis corresponds to the percentage of tweets that contains the term.*)

However, for the sake of developing gender identity questions on intake forms we want to know whether there are reasons to suspect differences among terms used by trans* people at the regional level. Furthermore, since the same intake form is used for both trans* people and non-trans* people, we also want to be able to compare the terminology used by the trans* community with the general population to minimize non-trans* patients inadvertently indicating a trans* status.[xiv] To gather data for distributional similarity measures, we performed the same term frequency and co-occurrence analysis for each state. Figure 2 compares the word clouds of the keywords

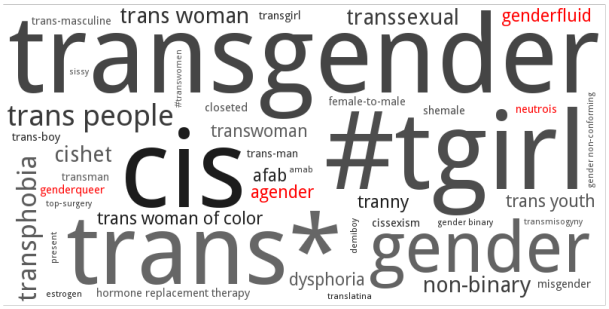
[xiv] For more on false negative transgender identification, we direct the reader to The GenIUSS Group. Gender-Related Measures Overview.

used in Arkansas between trans* people and the general public; while Figure 3 depicts the word clouds of the keywords used by trans* people in Arkansas, Florida, Washington, and Kansas. Consider the term ‘genderqueer’ which appears on all three of the sample intake forms in Table 1. An examination of the word clouds in Figure 3 shows that in Arkansas the general public uses the term ‘genderqueer’ (0.46%) more frequently than the term ‘genderfluid’ (0.23%). In contrast, the trans* population uses the term ‘genderfluid’ (0.57%) more often than the term ‘genderqueer’ (0.28%) in Arkansas. Similarly, as shown in Figure 3, the term ‘genderqueer’ while present in Kansas (0.17%) among trans* people, is used less frequently than ‘genderfluid’ (0.34%) and ‘agender’ (0.43%). In Washington, ‘genderqueer’ (0.77%) and ‘genderfluid’ (0.80%) are used with about the same relative frequency. However, in Florida, we see usage that is the inverse of Arkansas and Kansas; ‘genderqueer’ (0.43%) is used more frequently than ‘genderfluid’ (0.25%). Further, ‘agender’ is used with less relative frequency in Arkansas (0.28%) and Florida (0.33%) than in Kansas (0.43%) and Washington (0.46%).

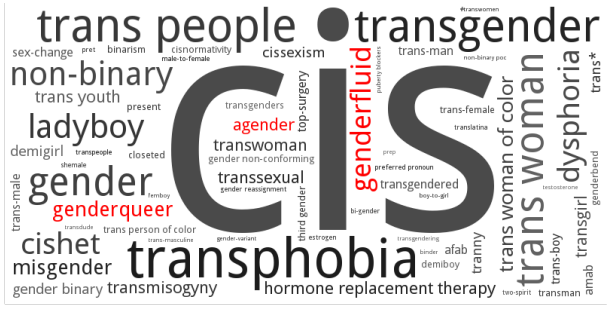

(a) AR: general public


(b) AR: trans* people

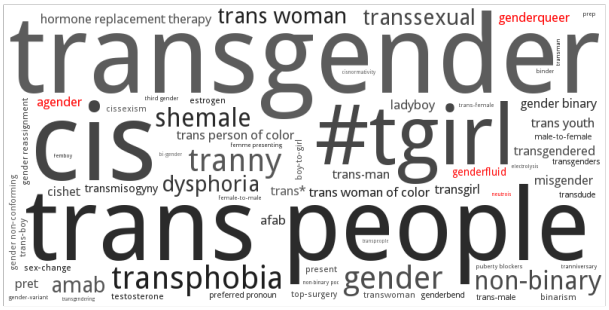
**Figure 2.** Word clouds of the keyterms in Arkansas: (a) trans* people vs. (b) the general public.




(a) Kansas




(b) Washington




(c) Florida

**Figure 3.** Word clouds for the keyterms used by trans* people in three states: (a) Kansas, (b) Washington, and (c) Florida.

Further, consider the term ‘neutrois’, which describes individuals who feel that they have no gender or are gender neutral. This is an example of a regionally specific term whose meaning cannot be characterized as “not exclusively male or female” (similar to ‘agender’), and as such would not be captured by the options of the sample intake forms surveyed. We found that users classified as trans* used the term ‘neutrois’ in only twelve states: CA, FL, GA, LA, MA, MI, MN, NY, PA, TX, VA, and WA. That is, in addition to states such as CA, MA, VA, and WA that are

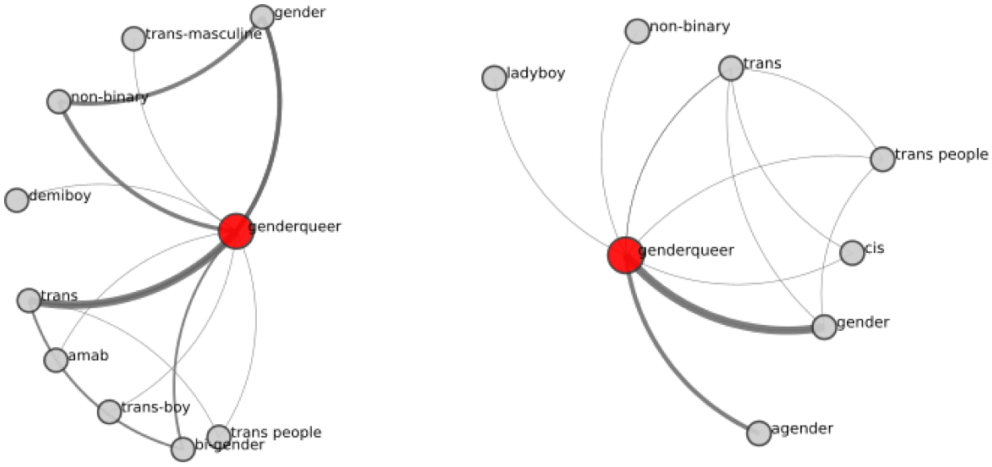
known for having a large identified LGBT population, ‘neutrois’ appears in the Great Lakes states and some of the Southern states.

**Discussion**

*Generating hypotheses about language preference*

While there are no definitive conclusions that can be drawn from this data alone, the findings suggest the following conjecture: intake forms in the southern United States that use ‘genderfluid’ and ‘agender’ rather than ‘genderqueer’ may have better response rates than forms that use ‘genderqueer’.

Furthermore, in light of the findings of the term ‘neutrois’ in the Great Lake states and select southern states, another conjecture is that trans* populations in these states have a higher incidence of ‘neutrois’ in the free-text ‘please specify’ fields.




(a) Washington (b) Florida

**Figure 4.** Comparing co-occurrence networks of the keyterms around ‘genderqueer’ used by trans* people in (a) Washington and (b) Florida.

Similar conjectures can be generated by comparing the co-occurrence networks at the individual state-level. For example, Figure 4 shows the structures of the two co-occurrence networks around ‘genderqueer’ for trans* in Washington and trans* in Florida. These co-occurrence networks show regional variations in the co-occurrence of ‘genderqueer’. For example, in Washington ‘genderqueer’ co-occurs with terms ‘trans-boy’ and ‘demi-boy’ that fit the Fenway Health characterization of ‘genderqueer’ (“not exclusively male or female”), but also terms such as ‘agender’ and ‘non-binary’ which are gender identities that do not fit this characterization. On the other hand, in Florida, ‘genderqueer’ does not co-occur with terms favoring one end of a binary spectrum, but it does co-occur with ‘non-binary’. From these observations, we can generate the following conjectures: ‘genderqueer’ denotes a broader set of gender-identities in Washington than in Florida, and in both cases it denotes identities not adequately characterized by Fenway Health’s gloss.

These conjectures, however, stand in need of further testing using formal and controlled methods. One of the trouble spots for our research is that some states have very few relevant tweets collected and few users classified as trans*. Delaware, Montana, and Wyoming each only had one user classified as trans* while South Dakota and Mississippi each has only two. While it is likely that this is because there are relatively few trans* persons in these regions using Twitter to discuss issues related to gender identity, it is also possible that trans* related tweets are not captured in our data set because the language used to discuss these terms are not in our list of keywords. We reviewed the raw data captured to date with the current keyterms to find additional keyterms we have missed, but did not find any in this set. These gaps in data point to the need for tools and methods outside of those discussed in this paper for gathering data and testing hypotheses about variations in transgender identity terms.

*Limitations*

Our study suggests that social media data sources such as Twitter can expand the range of what can be easily measured and provide new types of information for mining health-related knowledge. However, in addition to big

data challenges, Twitter data has its limitations and may not be reliable for answering certain questions. First, although Twitter has a set of feature-rich APIs and a relatively open policy for scraping, collecting relevant data to answer a specific scientific question is not easy. We collected over 154 million raw tweets in less than two months; however, only a fraction of the data (500,000 tweets) was deemed relevant to our study. Second, we found that even with a list of well-developed search terms, the returned data set had many false positives, which affirms the necessity of building classifiers to further narrow the search results. Nevertheless, the process of building classifiers is a tedious process involving manually annotating a large number of tweets to produce a gold-standard training data; and the accuracies of the classifiers were not perfect. In particular, the recall of the second classifier – finding self-identified trans* people – is low (0.261) indicating that we have missed many true positive cases. This might be the reason that we do not have a large enough corpus for trans* people. More sophisticated features[23,24] can be incorporated into the classifiers to improve the performance. Third, the geographic analysis was coarse-grained, providing only statistics on the state level. Although we attempted to geotag tweets with more fine-grained location information at the city level, the result was not satisfactory due to common conflicts in city names (e.g., Springfield, SC vs. Springfield, MA vs. Springfield, IL). Even though Twitter added the capability to record geocodes (latitude and longitude) and introduced new geographic metadata ('geo' and 'place'), there are very few tweets and user profiles we collected with geocodes available. One possible reason for this phenomenon is Twitter users having to give explicit consent to allow software vendors to record their geocodes. Another possible reason is that geocodes are only available if the tweets are sent from devices that have Global Positioning System (GPS) enabled. More sophisticated geocoding techniques[7,15] may be utilized to provide more accurate and finer grained location information. However, there is no direct way to integrate these techniques into our pipeline.

Finally, we note that our study is limited by the user demographics available on social media platforms. The users of social media tend to be younger (e.g., 37% of Twitter users are under 30, while only 10% are 65 or older, as of 2014[xv]); and there are power users who exhibit a substantially greater quantity of activity than the average user.[xvi] These characteristics are likely to create sample bias and impose limitations on mining meaningful information that represents a broader population. For instance, Twitter data may not be reliable for mining information about senior citizens.

**Conclusion**

This research shows that mining information on social media platforms such as Twitter can yield valuable insights to guide hypothesis generation in the development of intake questionnaires. While the output of this pilot study is insufficient to guide the development of better intake forms, it can be used to generate hypotheses for further testing. Furthermore, this data set can form the basis of future research in transgender health care. By capturing terms in context, we have generated a data set that will allow us to look at contextually sensitive aspects of the term use such as sentiment analysis in future research. Utilizing social networking resources also produces a data set that will allow us to begin investigating the social influences related health concerns among transgender persons, which is one of the priority research areas identified by the IOM. Finally, our experiences with mining Twitter data in this study yield a good process in dealing with large textual social media datasets.

**References**


1. Institute of Medicine. The health of lesbian, gay, bisexual, and transgender people: building a foundation for better understanding. Washington, DC: National Academy of Sciences. Washington, DC: National Academy of Sciences (US); 2011.
2. Catania JA., Binson D, Canchola J, Pollack LM, Hauck W, Coates TJ. Effects of interviewer gender, interviewer choice, and item wording on responses to questions concerning sexual behavior. *Public Opinion Quarterly* 1996; 60(3): 345–75.
3. Kuper LE, Nussbaum R, Mustanski B. Exploring the diversity of gender and sexual orientation identities in an online sample of transgender individuals. *Journal of Sex Research* 2012; 49(2-3); 244-54.
4. The GenIUSS Group. Gender-Related Measures Overview. Los Angeles (CA): The Williams Institute (US); 2013 Feb.
5. Hecht B, Hong L, Suh B, Chi EH. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. *Proc SIGCHI Conference on Human Factors in Computing Systems* 2011: 237-46.


[xv] http://www.pewinternet.org/2015/01/09/demographics-of-key-social-networking-platforms-2/
[xvi] http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/

6. Cheng Z, Caverlee J, Lee K. You are where you tweet: a content-based approach to geo-locating twitter users. *Proc 19th ACM International Conference on Information and Knowledge Management* 2010: pages759-768.
7. Mahmud J, Nichols J, Drews C. Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology* 2014; 5(3),): 47.
8. Salton G, Fox EA, Wu H. Extended Boolean information retrieval. *Communications of the ACM* 1983; 26(11): 1022-36.
9. Breiman L. Random forests. Machine *L*earning 2001; 45(1),): 5-32.
10. Kroeger PR. Analyzing grammar: an introduction. Cambridge University Press, 2005.
11. Lund K and Burgess C. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers* 1996, 28(2),): 203-8.
12. Bullinaria JA, Levy JP. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods* 2007; 39(3),): 510-26.
13. W3C SVG Working Group. Scalable Vector Graphics (SVG) 1.1 (Second Edition). [Online]. Available from: http://www.w3.org/TR/2011/REC-SVG11-20110816/. [Accessed March 1, 2015].
14. Bostock M, Ogievetsky V, Heer J. D3: Data-driven documents. *IEEE Transactions on Visualization & Computer Graphics* 2011; 17(12): 2301-9.
15. Eisenstein J, O'Connor B, Smith NA, Xing EP. A latent variable model for geographic lexical variation. *Proc Conference on Empirical Methods in Natural Language Processing* 2010: 1277-87.
16. Cahill S., Makadon H. Sexual orientation and gender identity data collection in clinical settings and in electronic health records: A key to ending LGBT health disparities. *LGBT Health*, 2014 1(1): 34-41.
17. Dargie E., Blair KL., Pukall CF., Coyle SM. Somewhere under the rainbow: Exploring the identities and experiences of trans persons. *The Canadian Journal of Human Sexuality* 2014; 23(2): 60-74.
18. Scheim AI., Bauer GR. Sex and gender diversity among transgender persons in Ontario, Canada: Results from a respondent-driven sampling survey. *Journal of Sex Research* 2015; 52(1): 1-14.
19. Carver CM. American regional dialects: A word geography. University of Michigan Press; 1989.
20. Chambers JK. Region and language variation. *English World-Wide*, 2001; 21(2):169-199.
21. Nerbonne J. How much does geography influence language variation? *Space in Language and Linguistics: Geographical, Interactional, and Cognitive Perspectives* 2013: 220-36.
22. Gouws S., Metzler D., Cai C., Hovy E. Contextual bearing on linguistic variation in social media. *Proc Workshop on Languages in Social Media 2011*: 20-29.
23. Mikolov T., Chen K., Corrado G., Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint* 2013; arXiv:1301.3781.
24. Bloehdorn S., Hotho A. Boosting for text classification with semantic features. *Proc WebKDD* 2004; 149-166.
25. Zwicky A. Two lavender issues for linguists. *Queerly phrased: Language, gender, and sexuality* .1997: 21-34.
26. Murphy ML. The elusive bisexual: Social categorization and lexico-semantic change. *Queerly phrased: Language, gender, and sexuality.* 1997: 35-57.
27. Lee L. Measures of distributional similarity *Proc the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* 1999: 25-32.