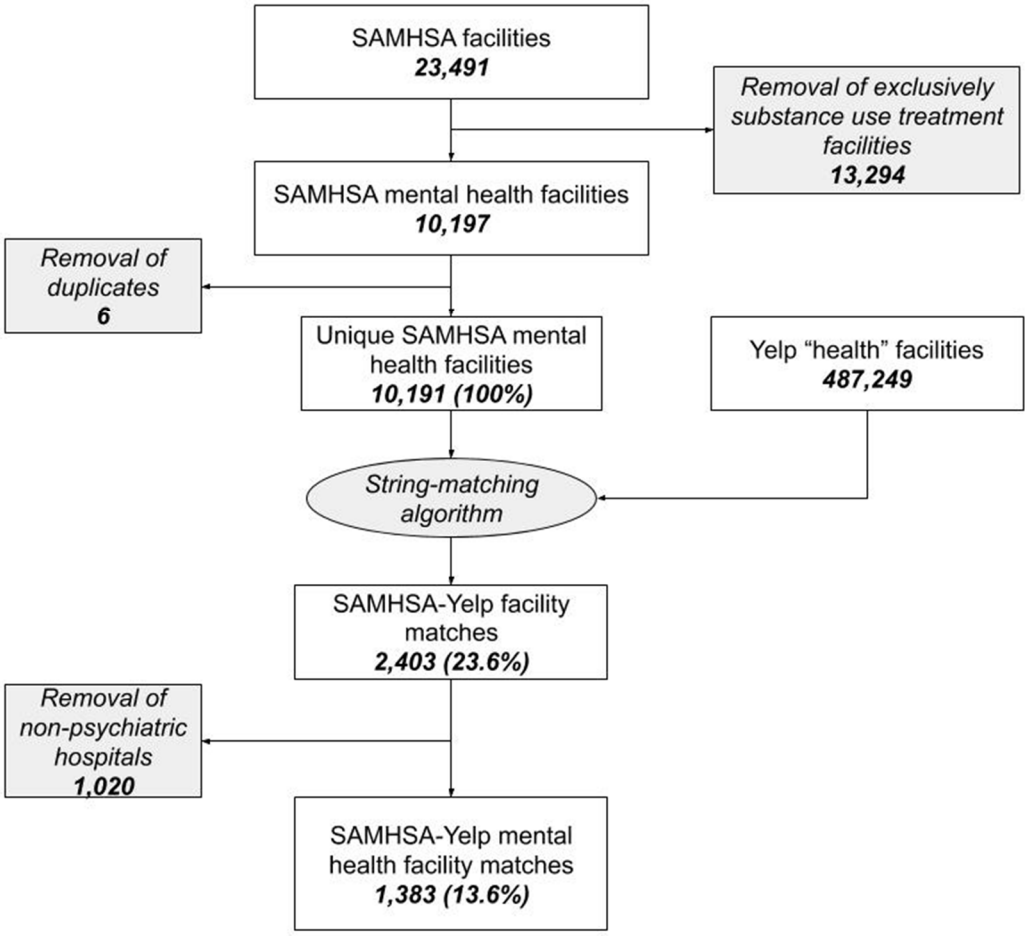
Online supplement for Stokes et al., DOI: 10.1176/appi.ps.202000267

Final dataset derivation

Flowchart showing derivation of the final dataset of SAMHSA-Yelp matched facilities



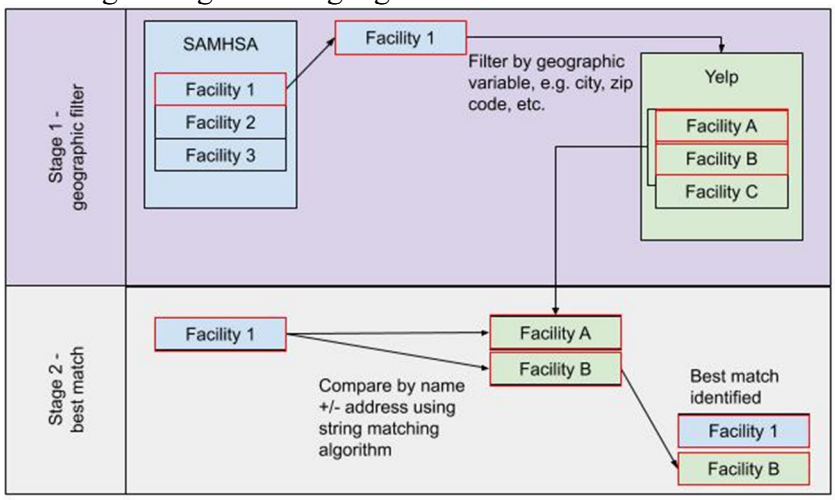
Development and testing of a string-matching algorithm

We developed a string-matching algorithm to match each SAMHSA facility to its corresponding Yelp page, where a Yelp page existed. We elected to match facilities through an algorithmic approach, rather than through hand matching, for two reasons. First, the volume of SAMHSA facilities (10,191) and of Yelp facilities (487,249), made hand matching prohibitively labor intensive. Second, algorithmic approaches allow for greater reproducibility. Both SAMHSA's Treatment Locator and Yelp are continuously updated, and a string-matching algorithm allows for replication of our approach even as the datasets continue to grow.

First, a random selection of 200 SAMHSA facilities (0.85%) were manually matched to Yelp facilities where possible. Several two-stage string-matching algorithms were then compared against this same sample. The best-performing algorithm was applied to the entirety of the SAMHSA data set.[1]

The first stage of the algorithm was treated as a geographic screening stage to identify a subset of Yelp facilities to compare with each SAMHSA facility. The second stage then compared SAMHSA facility names and addresses to the subset of Yelp facility names and addresses. The match with the greatest similarity in these two fields, with similarity defined by several different string-matching algorithms, was then selected as the most likely match for the given SAMHSA facility. Where the same Yelp facility was selected as the best match for multiple SAMHSA facilities, only the match with the greatest similarity was kept. A schematic representation of this process is shown below.

Two stage string-matching algorithm flow

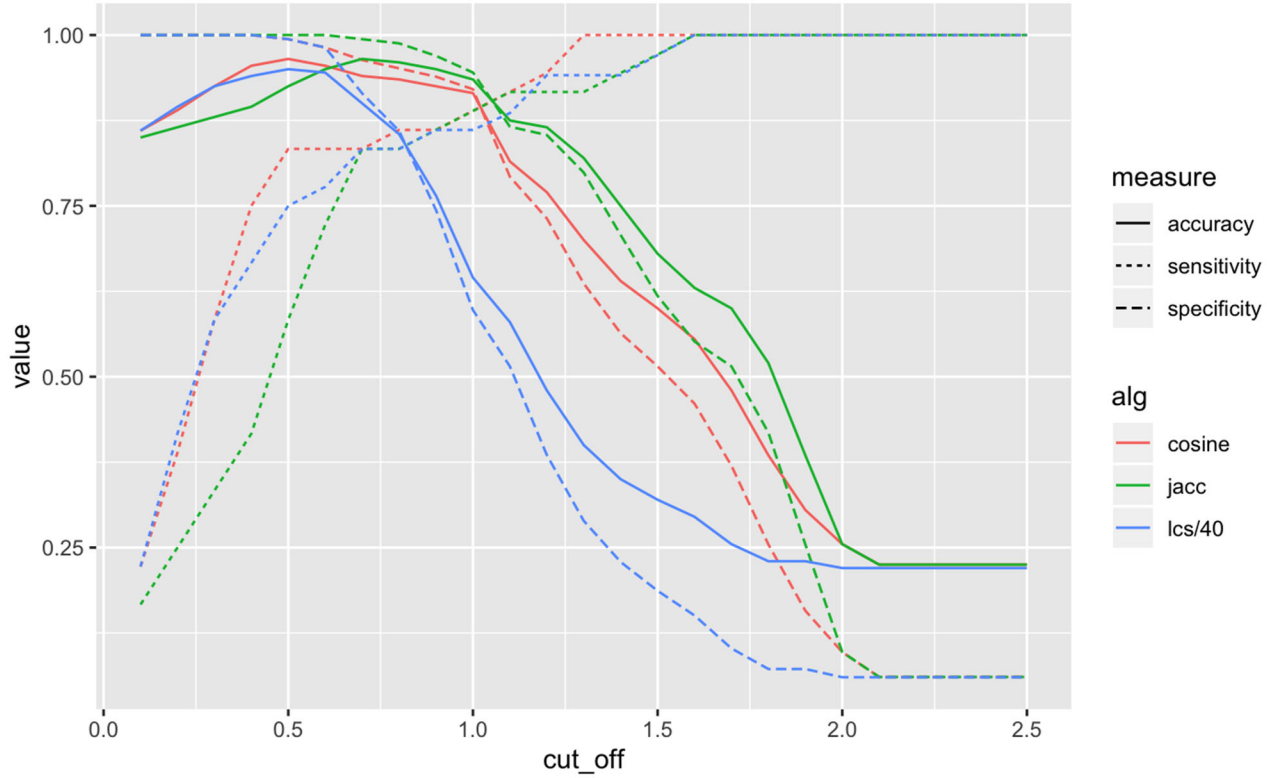


The variables tested in the first stage were “zip code”, “city”, and “address”. The variables tested in the second stage were “name” and “address”. In the second stage, three string matching algorithms were tested: cosine score, jaccard score, and longest common substring. Given that the facility names could include extra words, misspelled words, and altered word orders, qgram algorithms like jaccard score and cosine score were expected to perform better than replacement algorithms like longest common substring.[2]

In order to determine the best algorithm for facility name comparison, the subset of 200 hand coded facilities was tested with a first stage filter of (same zip) OR (same city AND same state) and a second stage string match of facility name and address for each of the three algorithms. For the qgram algorithms, a q of 4 was used for both name and address. The accuracy, sensitivity, and specificity were compared across the three algorithms for various cut-off scores. Jaccard and cosine scores were found to perform better than longest common substring, as expected. Both jaccard and cosine had the same maximum accuracy. As demonstrated in the figure below, cosine score had superior sensitivity while jaccard score had superior specificity. Sensitivity, or

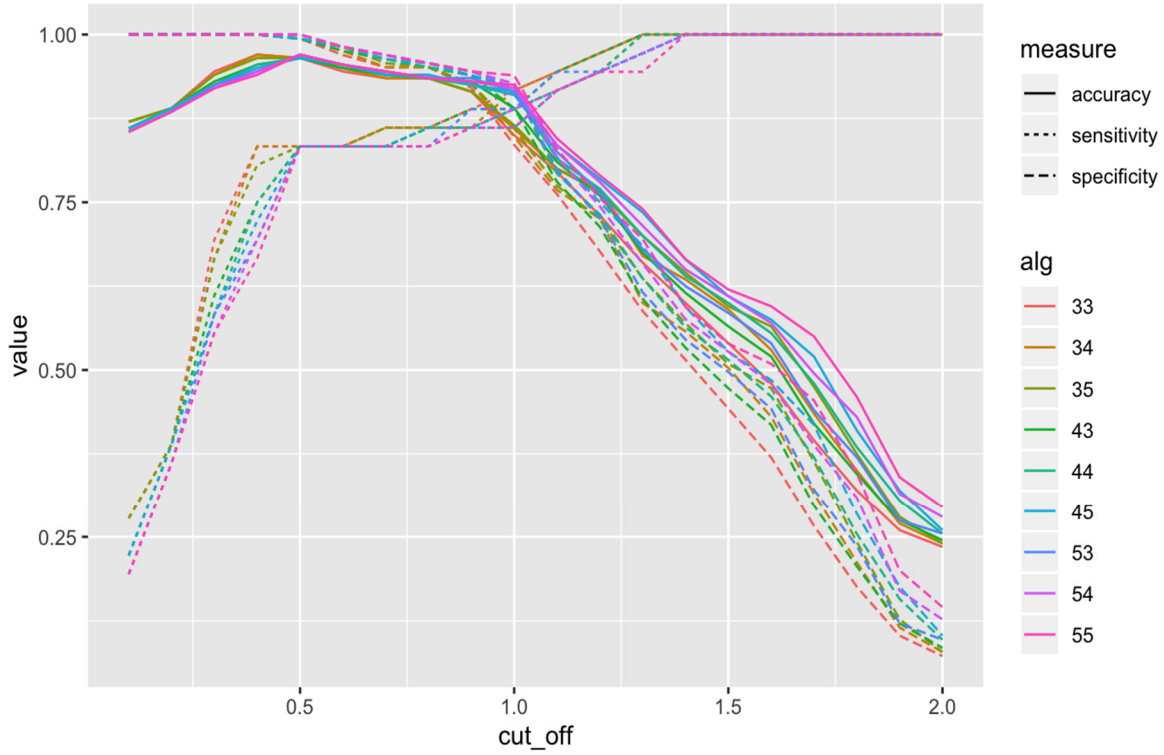
the ability to correctly identify true matches, was valued over specificity, or the ability to correctly identify the absence of a true match. Cosine score was therefore selected as the string-matching algorithm of choice.

Accuracy, sensitivity, and specificity of cosine score, jaccard score, and longest common substring score for sum of facility name and address comparison using various cut-off values for "true match"



Once cosine score was selected, we tested multiple iterations of qgram size for address and name comparisons. In general, smaller qgrams corresponded to higher sensitivity and lower specificity, whereas larger qgrams corresponded to lower sensitivity and higher specificity. Again, greater value was placed on optimizing sensitivity. Based on evaluation of the figure below, qgram values of 4 and 3 were selected for name and address, respectively.

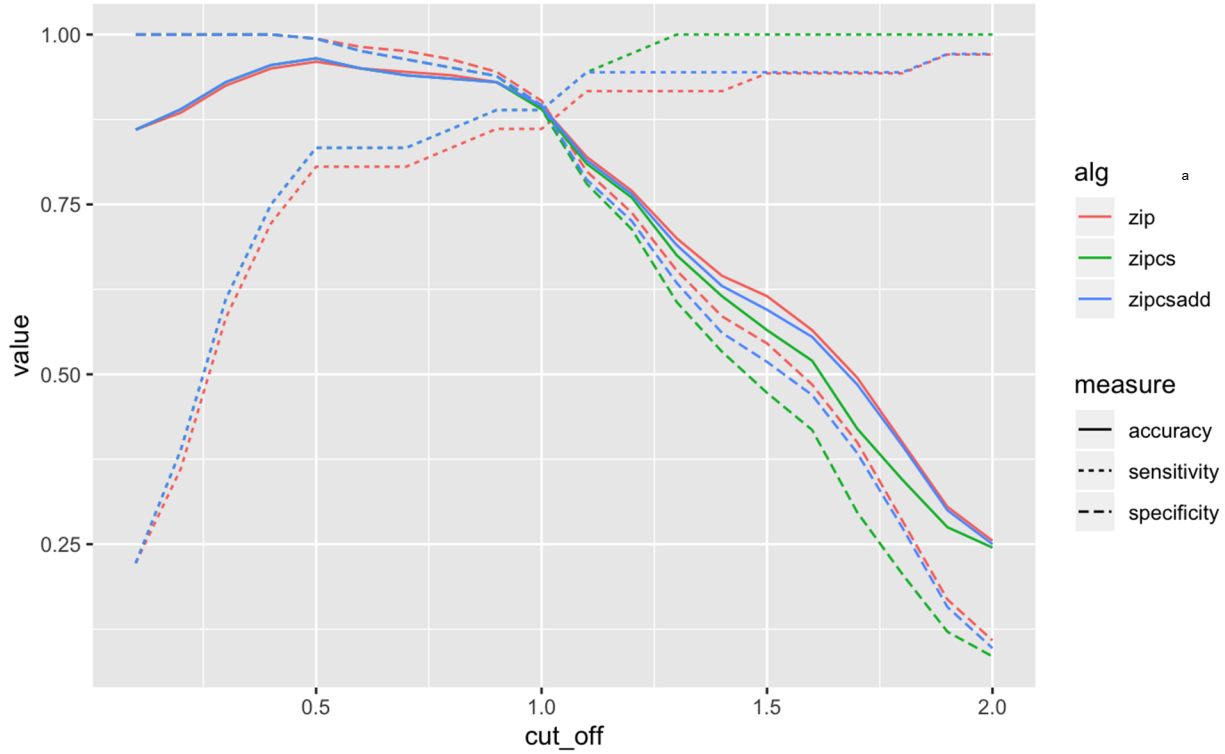
Accuracy, sensitivity, and specificity of various qgram values for facility name and address comparisons



Note: "alg" reflects the qgram for name, followed by the qgram for address

Finally, using the cosine score with these qgram values, the first step of the matching algorithm was modified to assess for the impact of solely using zip code, using zip code OR (city AND state) and using zip code OR (city AND state AND address similarity, defined by a longest common substring<5). Zip code OR (city AND state) had the highest sensitivity, as shown in eFigure 4, and was therefore selected.

Accuracy, sensitivity, and specificity of first stage filtering by zip; zip and city/state; or zip, city/state and address similarity



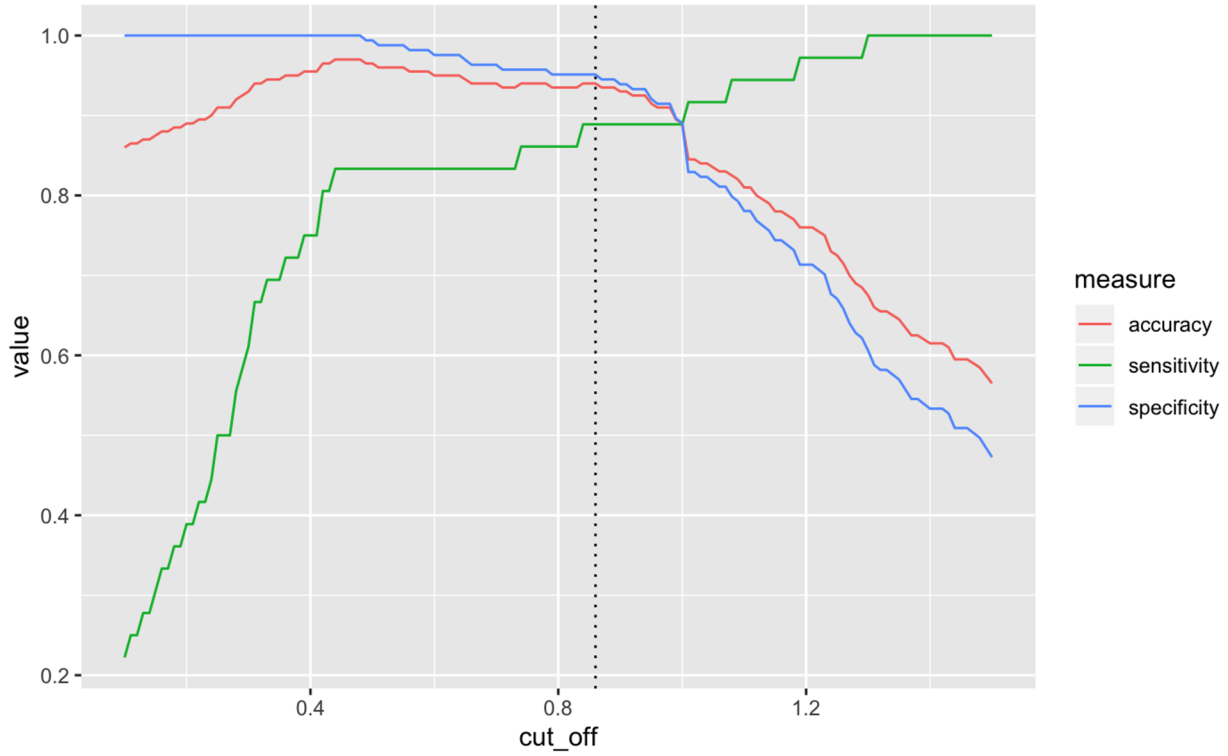
"zipcs" = zip code OR (city AND state); "zipcsadd" = zip code OR (city AND state AND address similarity, defined by a longest common substring <5)

The final algorithm, therefore, employed a first step selection of match on zip code or city/state and a second step selection of lowest sum of facility name cosine score (qgram of 4) and facility address cosine score (qgram of 3).

This algorithm was applied to the 23,482 SAMHSA facilities (both "mental health" and "substance use"). Where the same Yelp facility was matched to multiple SAMHSA facilities, only the match with the lowest composite cosine score was considered a true match. The final match results were compared to the 200 hand coded results to determine an optimal cut-off score. A cut-off value of 0.86 was selected based on an analysis of the changes in accuracy, sensitivity, and specificity as the cut-off was increased from 0.1 to 1.5 by 0.01. After 0.86, specificity and

accuracy suffered for further improvements in sensitivity, as shown below. The final algorithm had a sensitivity of 0.89 and a specificity of 0.95.

Accuracy, sensitivity, and specificity of the final algorithm with changes to the cut-off value for defining a "true match"



Strategies for the removal of Yelp pages pertaining to providers of general medical care

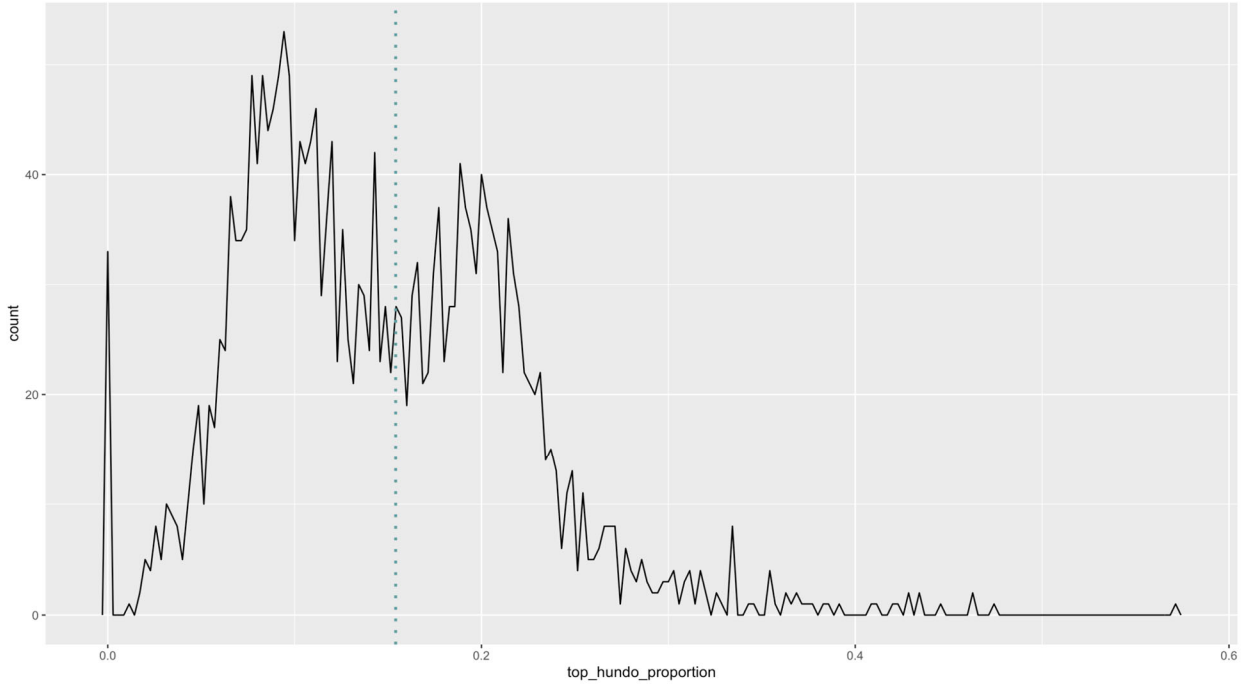
Following application of the string matching algorithm, 2,403 unique SAMHSA mental health facilities were matched with high probability to Yelp facilities (23.6 % of all SAMHSA mental health facilities). There were a total of 33,532 reviews associated with these facilities. These matches included hospitals and medical centers with psychiatric units. The reviews in these cases tended to correspond to the parent organization, and therefore general medical care, rather than mental health treatment services. The high number of reviews pertaining to general medical care affected latent Dirichlet allocation (LDA) topic definition, such that many more categories pertained to medical care than to psychiatric care.

We therefore attempted two strategies for the removal of hospitals and medical centers from the dataset. First, we matched the list of Hospital Compare hospitals to Yelp facilities using the string matching algorithm and cut-off previously described. Of the 5,334 Hospital Compare facilities, 3,618 (67.8%) matched to a Yelp facility. After removing non-psychiatric Hospital Compare facilities matched to a Yelp facility from the list of Yelp-matched SAMHSA facilities, 1,836 SAMHSA facilities remained with 12,380 corresponding reviews. LDA of these reviews showed some improvement in the proportion of psychiatric care LDA topics relative to medical care topics, but still a majority of topics pertaining to medical care.

Second, we undertook a manual review of a random sample of 211 facilities. We used facility websites to identify each facility as either a provider of general medical care or a provider of primarily mental health care. Of the 209 facilities, 94 were confirmed to be hospitals or medical centers, and 115 were confirmed to be primarily mental health providers. We identified the 100

words with the greatest difference in frequencies between the reviews of the medical centers and those of the mental health facilities. These words included “hospital”, “er”, “surgery”, “delivery”, and “blood”. We looked at the distribution of the frequency of use of these 100 words in the reviews for all Yelp-matched SAMHSA facilities. As shown in the figure below, the distribution was bimodal. We used Jenk’s natural breaks algorithm to identify a breakpoint of 0.154. We assumed that facilities whose reviews included a proportion of the 100 words above that breakpoint were hospitals (and should therefore be removed from the data-set), and that the remaining facilities were non-hospitals. This resulted in a final sample of 1,383 facilities with 8,133 reviews. An LDA of the remaining facilities showed a much greater proportion of psychiatric care topics relative to medical care topics. This approach was thus favored over the previously described removal of Hospital Compare hospitals. Applying the cut-off to the hand-coded sample showed a sensitivity of 0.94 and a specificity of 0.91 for identification of mental health treatment facilities.

Distribution of the percent of all words in reviews for a given facility among the top 100 words with greatest difference in frequency between confirmed non-psychiatric hospitals and primarily mental health facilities



Note: the dotted blue line shows the Jenk's natural breakpoint

References

1. McKenzie G, Janowicz K, Adams B. Weighted multi-attribute matching of user-generated points of interest. In: *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM; 2013:440–443.
2. Van der Loo MP. The stringdist package for approximate string matching. *R J*. 2014;6(1):111–122.