JAMIA Open, 4(2), 2021, 1–9
doi: 10.1093/jamiaopen/ooab042
Research and Applications

Research and Applications

Automatic gender detection in Twitter profiles for health-related cohort studies

Yuan-Chi Yang[1], Mohammed Ali Al-Garadi[1], Jennifer S. Love[2], Jeanmarie Perrone[3], and Abeed Sarker[1,4]

[1]Department of Biomedical Informatics, School of Medicine, Emory University, Atlanta, Georgia, USA[2]Department of Emergency Medicine, School of Medicine, Oregon Health & Science University, Portland, Oregon, USA[3]Department of Emergency Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA and [4]Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, Georgia, USA

Corresponding Author: Yuan-Chi Yang, PhD, Department of Biomedical Informatics, School of Medicine, Emory University, 101 Woodruff Circle, 4th Floor East, Atlanta, GA 30322, USA; yuan-chi.yang@emory.edu

Received 21 January 2021; Revised 27 April 2021; Editorial Decision 3 May 2021; Accepted 4 May 2021

ABSTRACT

Objective: Biomedical research involving social media data is gradually moving from population-level to targeted, cohort-level data analysis. Though crucial for biomedical studies, social media user's demographic information (eg, gender) is often not explicitly known from profiles. Here, we present an automatic gender classification system for social media and we illustrate how gender information can be incorporated into a social media-based health-related study.

Materials and Methods: We used a large Twitter dataset composed of public, gender-labeled users (Dataset-1) for training and evaluating the gender detection pipeline. We experimented with machine learning algorithms including support vector machines (SVMs) and deep-learning models, and public packages including M3. We considered users' information including profile and tweets for classification. We also developed a meta-classifier ensemble that strategically uses the predicted scores from the classifiers. We then applied the best-performing pipeline to Twitter users who have self-reported nonmedical use of prescription medications (Dataset-2) to assess the system's utility.

Results and Discussion: We collected 67 181 and 176 683 users for Dataset-1 and Dataset-2, respectively. A meta-classifier involving SVM and M3 performed the best (Dataset-1 accuracy: 94.4% [95% confidence interval: 94.0–94.8%]; Dataset-2: 94.4% [95% confidence interval: 92.0–96.6%]). Including automatically classified information in the analyses of Dataset-2 revealed gender-specific trends—proportions of females closely resemble data from the National Survey of Drug Use and Health 2018 (tranquilizers: 0.50 vs 0.50; stimulants: 0.50 vs 0.45), and the overdose Emergency Room Visit due to Opioids by Nationwide Emergency Department Sample (pain relievers: 0.38 vs 0.37).

Conclusion: Our publicly available, automated gender detection pipeline may aid cohort-specific social media data analyses (https://bitbucket.org/sarkerlab/gender-detection-for-public).

Key words: natural language processing, machine learning, Twitter, user profiling, gender detection, toxicovigilance

LAY SUMMARY

To perform biomedical research using social media data on a targeted cohort, the user's demographic information (eg, gender) is typically required. However, the information is often not explicitly known from the user profile. One solution is to infer the information from the user's public data via natural language processing and machine-learning techniques. In this work, we focused on estimating the user's gender and developed a highly accurate pipeline. We then applied the pipeline on a Toxicovigilance cohort of Twitter users who have self-reported misuse of prescription medications (PMs), including tranquilizers, stimulants, and opioids. We found that the pipeline performs with high accuracy on this data set. Additionally, the inferred gender proportions of those users are consistent with traditional surveys, including the National Drug Use and Health Survey 2018 by the Substance Abuse and Mental Health Services Administration and the estimated overdose-related Emergency Department visits in 2016 from the Nationwide Emergency Department Sample. The results support that social media data can be harnessed as a complementary source to traditional surveys and can be used to understand the demographics of PM misuse in the United States. Our gender detection pipeline will be made publicly available to ensure transparency and support community-driven development.

INTRODUCTION

Social media data are increasingly being used for health-related research.[1,2] Users often discuss personal experiences or opinions regarding a variety of health topics, such as health services or medications.[1–3] Such information can be categorized, aggregated and analyzed to obtain population-level insights,[4–8] at low cost and in close to real time. It has thus been used as a resource for population health tasks such as influenza surveillance, pharmacovigilance, and toxicovigilance.[9–11] While early research mostly attempted to conduct observational studies on entire populations (eg, Twitter users discussing flu),[12] some recent studies have been moving to targeted cohorts (eg, pregnant women,[13] people in certain geo-locations,[14] cancer patients,[15] and people suffering from mental health issues[16–19]). Demographic information about such cohorts can help researchers investigate what roles demographics have in a given study, understand if social media is biased toward specific cohorts, and explicitly address these biases.[20,21] Due to the importance of explicitly considering biological sex or gender in health research, funding agencies, including the National Institutes of Health, have emphasized the necessity to describe sex/gender information of the cohorts included in research studies (eg, through inclusion of women).[22] This, however, presents a challenge for social media-based studies because the demographic information of the users are often not explicitly known.

One solution is to infer the demographic information from the users' metadata. In the past two decades, researchers have developed various automatic methods for characterizing users. Taking gender detection on Twitter as an example, researchers have investigated classification schemes based on the users' (screen) names, profile descriptions, tweets, profile colors, and even images, with machine learning algorithms such as support vector machine (SVM), Naive Bayes, Decision Tree, Deep Neural Network, and Bidirectional Encoder Representations from Transformers (BERT).[23–33] Some have made their pipelines publicly available and have since been applied to social media mining tasks. For example, Sap et al[26] released a lexicon for gender and age detection and it was applied for mental health research.[16–18] Knowles et al[28] released a package named Demographer to infer gender based on users' first names and it was later employed to infer gender in studies for influenza vaccination[34] and mental health.[19] Wang et al[31] also released a multimodal deep learning system (M3) to infer gender based on users' profile information, including pictures, (screen) names, and descriptions. Though these existing pipelines can be directly applied to biomedical tasks, there is still room for improvement, particularly for Twitter data. First, none of these pipelines used all four of the users' textual attributes—names, screen names, descriptions, and tweets. This is a missed opportunity and there is thus the possibility to further improve upon these models by developing a pipeline capable of incorporating these four attributes or more. Second, these experiments have not been validated on the same data, making it impossible to perform direct comparisons of their performances. Third, to the best of our knowledge, these pipelines were developed based on general users, but have not been tested on gender-labeled, domain-specific datasets. Benchmarking the performance variations due to domain change can inform researchers about the applicability of these pipelines on their specific tasks.

In this work, we aimed at developing a high-accuracy, automatic gender classification system and evaluated its performance and utility on a domain-specific dataset. In the following sections, we first describe our experiments with various unimodal and multimodal strategies and existing pipelines, and compare their performances on a unified platform. We then discuss the benchmarking of the best strategies on our domain-specific (Toxicovigilance) dataset, consisting of a Twitter cohort of self-reported nonmedical consumers of prescription medications (PMs). The benchmarking involves evaluating performance scores on an annotated subset. To illustrate the utility of this pipeline, we applied the best-performing approach to compare the inferred gender proportions of a Twitter cohort with traditional, trusted sources.[35,36] The source code for gender detection experiments described will be made open source (https://bitbucket.org/sarkerlab/gender-detection-for-public).

MATERIALS AND METHODS

This study was approved by the Emory University institutional review board (IRB00114235).

Gender detection pipeline development

Data collection

We collected gender-labeled datasets for general Twitter users, released by previous work.[25,33] The data from Liu and Ruths[25] consists of 12 681 users with binary annotations obtained via crowdsourcing through Amazon Mechanical Turk.[37] Each instance was coded by three annotators and a label was accepted only if all three annotators agreed. The data from Volkova et al[33] consists of 1 000 000 tweets, randomly sampled from the data in Burger et al,[23]

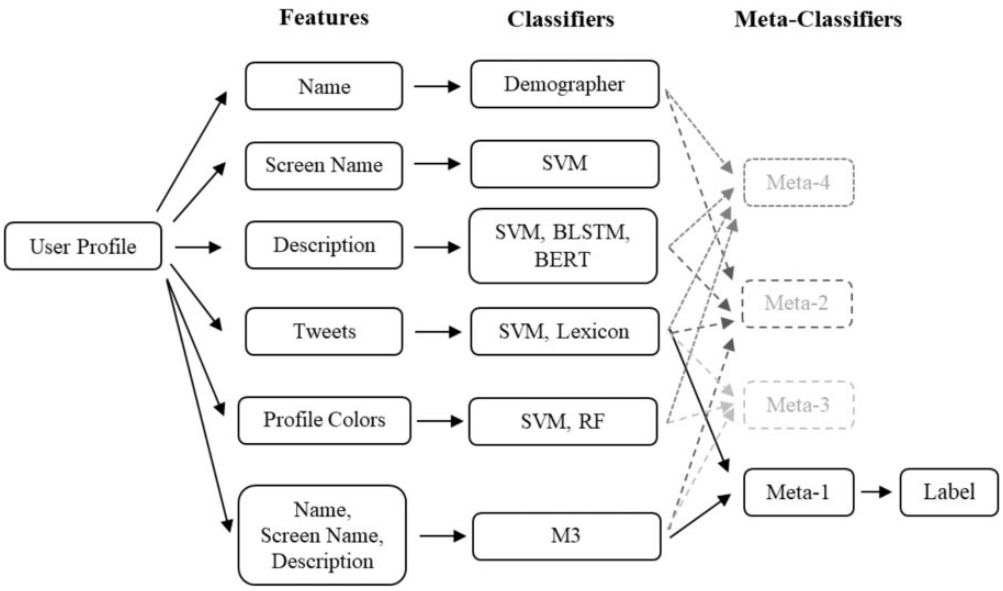

Figure 1. Gender classification pipeline, from user profile to gender label.

which is labeled using users' self-specified genders on Facebook or MySpace profiles linked to their Twitter accounts. Both datasets provide the users' IDs and gender labels. Our focus is to develop the informatics infrastructure to detect gender as Twitter users self-identify themselves on the social media platform and we consider the two annotation methods to fall within this definition. We combined the two datasets and extracted users' publicly available data using Twitter API, including profile meta-data, such as handle names, descriptions, and profile colors, as well as the users' timelines (only English tweets were collected, while the retweets were excluded; users who had no original English tweets were dropped). We called this dataset as Dataset-1 and split it into training (60%), validation (20%), and test (20%) sets for pipeline development.

Classification

We first developed classifiers based on single attributes (ie, unimodal), including names and screen names, descriptions, tweets, and profile colors. We then experimented with building meta-classifiers based on the predicted scores from these classifiers (ie, multimodal). The flowchart in Figure 1 illustrates our processing pipeline. In the experiments, we considered machine learning algorithms including SVMs,[38,39] Random Forest (RF),[40] Bi-directional Long Short-Term Memory (BLSTM),[41,42] and BERT,[43,44] as well as existing resources including the lexica released by Sap et al,[26] the Demographer system by Knowles et al[28] and the M3 system (without profile picture) by Wang et al[31] Below we briefly outline each experiment, with further details in the Supplementary Table S1.

Name and screen name. We applied package Demographer[28] (DG) on the users' names. DG attempts to identify gender using character n-grams of user's first name, trained using the list of given names from US Social Security data. Similar to DG, we trained an SVM classifier for screen names using character n-grams.

Description. To classify gender using a user's description, we experimented with SVM, BLSTM, and BERT, approaches suited for free text data. BERT is a transformer-based model that produces contextual vector representations of words and achieves state-of-the-art performance on many tasks.[43,45] Many models with similar architecture have then been implemented and released.[46,47]

Each description was pre-processed by lowercasing and anonymizing URLs and user names. For SVM, the features are the normalized term frequency of unigrams. For BLSTM and BERT, each word or character sequence was replaced with a dense vector, and the vectors were fed into the algorithms for training.

Tweets. Focusing on users who have a substantial number of tweets, we selected users in the training data with at least 100 tweets and merged all collected tweets as the training texts for experiments on SVMs. The pre-processing is the same as that for the SVM classifier using description. The regularization parameter was optimized according to the validation accuracy.

Colors. We utilized 5 features associated with profile colors, including background color, link color, sidebar border color, sidebar fill color, and text color. Each profile color is represented using RGB values, each ranging from 0 to 255. We collapsed each value into 4 groups, yielding 64 groups for each color. We then experimented with SVM and RF.

Meta-classifier. We experimented with building SVM models on the predicted scores from 4 different combinations of the classifiers:

- meta-1: SVM on tweets and M3.
- meta-2: SVM on tweets, M3, Demographer on name, and BERT on description.
- meta-3: SVM on tweets, M3, and SVM on colors.
- meta-4: DM on names, SVM on screen names, BERT on description, and SVM on tweets.

Classification performance evaluation. The classification performance evaluation is based on class-specific precision, recall, and F_1 score, as well as accuracy (male and female combined). These metrics are defined as the follows:

$$\text{precision} = \frac{\text{numberoftruepositiveinstances}}{\text{numberofpositiveinstances}}$$

$$\text{recall} = \frac{\text{numberoftruepositiveinstances}}{\text{numberofrelevantinstances}}$$

$$\text{F}_1\text{score} = \frac{2}{1/\text{precision} + 1/\text{recall}}$$

$$\text{accuracy} = \frac{\text{numberofcorrectlyclassifiedinstances}}{\text{numberofinstances}}$$

where F_1 score is the harmonic mean of precision and recall. We also calculate the area under the receiver operating characteristic curve (AUROC). The receiver operating characteristic curve presents the relationship between the true positive rate and the false positive rate under different threshold and the AUROC provides a measure for the performance. The range of AUROC is from 0 to 1, with 1 being the best.

Coverage. Some users have missing profile information such as name or description or use non-English characters in the name field. This may make the inference using the specific information impossible. Therefore, for each classifier, we show the percentage of users whose genders can be inferred from the relevant profile information (as "coverage") while the performance is evaluated using this subset of users.

Application on Toxicovigilance dataset

Data collection

To conduct Toxicovigilance research using social media, we had collected publicly available, English tweets mentioning over 20 PMs that have the potential for nonmedical use or misuse. The lists of PMs can be found in Supplementary Table S2. In our prior work, we have developed annotation guidelines with our domain expert (JP) and have annotated a subset consisting of 16 433 tweets.[48] A brief description of annotation guideline and example tweets are given in Supplementary Tables S3 and S4, respectively. Based on this set, we then developed automatic classification schemes to detect if the tweets are describing self-reported nonmedical use (referred as "misuse tweets" in the following).[49] In this work, we used this classifier to classify a dataset collected from March 6, 2018 to January 14, 2020 and extracted the users' publicly available data. We referred to this set as Dataset-2. We also grouped users whose misuse tweets could be geo-located in the United States as a subset (Dataset-2-US).[50]

Since Dataset-2 did not have manual binary annotations, we relied on a secondary source to identify a user's gender—their self-identified gender information on the linked public Facebook profiles—whenever possible. These users make up the test set of Dataset-2 for benchmarking.

Classification performance

We applied the best-performing classification strategies on the test set of Dataset-2 to evaluate their performances. This serves not only as a benchmarking of how those pipelines perform on the Toxicovigilance dataset (Dataset-2) but also provides a measure of the transferability of our pipelines across research problems.

Table 1. Data distributions for the training, validation and test sets from Dataset-1

Dataset	F	M	Total
Training (Dataset-1)	21 521	18 788	40 309
Validation (Dataset-1)	7133	6303	13 436
Test (Dataset-1)	7158	6278	13 436
Total (Dataset-1)	35 812	31 369	67 181

Gender distribution inference

To assess the utility of our cohort characterization pipeline on a health surveillance related task, we applied the best-performing classification pipeline on Dataset-2 (and Dataset-2-US) and analyzed the gender distributions of the users who had self-reported misuse/abuse on one of the three abuse-prone PM categories—stimulants (eg, Adderall®), which can increase alertness, attention, and energy and are mostly prescribed to treat Attention Deficit Hyperactivity Disorder, tranquilizers (eg, alprazolam/Xanax®), which slow brain activity and are mostly used to treat anxiety, and pain relievers (eg, Oxycodone/OxyContin®), specifically for those containing opioids.[35,51] We then compared the distributions with metrics from the 2018 NSDUH,[35] as well as the overdose-related Emergency Department Visits (EDV) in 2016 from the Nationwide Emergency Department Sample (NEDS).[36] The details of the calculation are given in the Supplementary Materials. We performed Pearson's Chi-squared test for contingency table to determine if the differences in the proportions of females inferred from the different sources (Twitter vs survey data) are statistically significant, defined as P-value < 0.05.

RESULTS

Gender detection pipeline development

Data Collection (Dataset-1)

In total, we were able to retrieve the user data from 67 181 users, consisting of 35 812 (53.3%) females (F) and 31 369 (46.7%) males (M), which is close to the distribution estimated by Burger et al[23] and Heil and Piskorski[52] (55% female and 45% male) but deviate from the distribution estimated by Liu and Ruths[25] (65% female and 35% male). The distribution is presented in Table 1.

Classification

The performance (F_1-score, accuracy, and AUROC) for each classifier and meta-classifier is presented in Table 2, while the precisions and recalls are presented in the Supplementary Table S5. We now highlight the main findings.

The best performing classification schemes were the meta-classifiers using predicted scores from M3 and SVM on tweets (meta-1, 2, 3), with accuracies around 94.4%. The second best scheme was meta-4, with an accuracy of 92.5%. These all performed better than existing pipelines, including the lexicon (86.5%), the Demographer (80.2%), and M3 (90.0%), and other unimodal classifiers.

Application on toxicovigilance dataset

Data Collection (Dataset-2)

We were able to retrieve past data from 176 683 users for Dataset-2 (63 306 users for Dataset-2-US). Less than 0.3% of the users (412) had publicly available gender information from linked Facebook

Table 2. Test results (on Dataset-1) for classifiers and meta-classifiers

Feature/method	F_1 score (95% CI) (0.XXX)		Coverage (%)	Accuracy (95% CI) (%)	AUROC
	F	M			
Name/DG	802 (795–810)	802 (795–809)	98.1	80.2 (79.5–80.9)	0.878
Screen name/SVM	748 (740–756)	719 (710–728)	100.0	73.4 (727–742)	0.817
Description/SVM	728 (719–736)	693 (683–703)	88.9	71.1 (70.3–72.0)	0.796
Description/BLSTM	724 (716–733)	665 (655–675)	88.9	69.7 (68.9–70.6)	0.781
Description/BERT	790 (782–797)	757 (748–766)	88.9	77.4 (76.7–78.2)	0.873
Tweets/SVM	893 (888–898)	879 (872–885)	100.0	88.6 (88.1–89.2)	0.933
Tweets/Lexicon	874 (868–880)	856 (849–862)	100.0	86.5 (86.0–87.1)	0.917
Profile/M3	903 (897–908)	898 (893–903)	100.0	90.0 (89.5–90.5)	0.968
Colors/SVM	671 (662–682)	649 (640–659)	100.0	66.1 (65.3–66.9)	0.712
Colors/RF	660 (651–669)	640 (630–649)	100.0	65.0 (64.2–65.8)	0.692
Meta-1	**947** (944–951)	**940** (936–944)	100.0	**94.4** (94.0–94.8)	**0.965**
Meta-2	**947** (943–951)	**939** (935–944)	100.0	**94.3** (93.9–94.7)	**0.971**
Meta-3	**948** (944–952)	**941** (937–945)	100.0	**94.5** (94.1–94.9)	**0.966**
Meta-4	930 (925–934)	920 (915–925)	100.0	92.5 (92.1–92.9)	0.953

Table 3. Test results (on Dataset-2, for users who have revealed gender information on Facebook) for classifiers and meta-classifiers

Feature/method	F_1 score (95% CI) (0.XXX)		Coverage (%)	Accuracy (95% CI) (%)	AUROC
	F	M			
Name/DG	717 (655–773)	833 (796–867)	94.9	79.0 (74.9–82.9)	0.844
Screen name/SVM	692 (634–745)	776 (732–816)	100.0	74.0 (69.7–78.2)	0.838
Description/BERT	674 (616–727)	715 (663–762)	94.9	69.6 (65.0–74.2)	0.839
Tweets/SVM	821 (772–865)	894 (864–921)	100.0	86.7 (83.3–89.8)	0.916
Tweets/Lexicon	770 (717–818)	846 (810–879)	100.0	81.6 (77.7–85.2)	0.889
Profile/M3	894 (855–928)	936 (913–956)	100.0	92.0 (89.3–94.4)	**0.974**
Meta-1	**927** (894–954)	**955** (935–972)	100.0	**94.4** (92.0–96.6)	**0.964**
Meta-4	885 (846–919)	926 (902–949)	100.0	91.0 (88.1–93.7)	0.955

profile pages. One hundred fifty-five out of 412 users in this subset were female (37.6%), while 257 users were male (62.4%).

Classification performance

The performances of the pipelines on the test set of Dataset-2 are shown on Table 3 (precisions and recalls are on Supplementary Table S6). The best performing pipeline was meta-1 (accuracy 94.4%). Besides M3 and meta-1, all the classifiers experience performance drops possibly due to domain change. Here, we left out meta-2 and meta-3 because meta-1 provides comparable performance while being simpler. We also note that the accuracy of meta-1 is 95.8% (95% confidence interval 93.3–98.3%) when restricted to users whose misuse tweets could be geo-located in the United States (239 users with 103 females and 136 males).

Gender distribution inference

We applied meta-1 on all the users and analyzed the gender distributions for the users who have self-reported abuse/misuse of tranquilizers, stimulants, or pain relievers (opioids). In Table 4, we report the number of users for each category, and the percentage of males and females, inferred through the classification results (meta-1), and reported by NSDUH 2018.[35]

Although the users in Dataset-2-US are only roughly one-third of all users in Dataset-2, the gender proportions are close to each other. For tranquilizer and stimulants users, the gender proportions inferred from Twitter are very close to the comparator from NSDUH 2018 (with no statistically significant difference for tranquilizer users). In contrast, the gender proportions of pain reliever users are quite different from the comparator from NSDUH 2018, but much closer to the overdose EDV from NEDS.[36] This suggests that Twitter data could be an indicator of the gender distribution of opioid overdoses and might provide complementary information to better understand the discrepancies between the aforementioned two traditional data sources.

DISCUSSION

Model performance and improvement

Meta-1 performs with high accuracy consistently across Dataset-1 (94.4%) and Dataset-2 (94.4%), better than all the existing pipelines and other classifiers on this platform. This shows building the gender detection pipeline based on the four prominent textual features (name, screen name, description, and tweets) can improve performance over existing approaches. Also, except meta-1 and M3, all classifiers performed worse on the domain-specific data. This illustrates the importance of benchmarking the existing machine learning systems on the targeted cohorts, in order to evaluate their applicability on the desired tasks. It also indicates that multimodal strategies could enhance the robustness of the system against unseen data and is thus desirable when building similar user-characterization pipelines.

Table 4. Gender distributions for selected medication categories (inferred by the classifier/NSDUH 2018/overdose EDV 2016)

Medication category	Number of users (geo-located in the US)	Percentage of male/female		
		inferred (geo-located in the US)	NSDUH 2018	overdose EDV 2016
Tranquilizers	62 471 (20 863)	0.499/0.501 (0.490/0.510)	0.499/0.501	—
Stimulants	93 598 (36,323)	0.504/0.496[a] (0.514/0.486[a])	0.551/0.449	—
Pain relievers	38,299 (12,077)	0.621/0.379[a] (0.635/0.365[a])	0.518/0.482[b]	0.630/0.370

[a]The female proportion whose difference with the corresponding female proportion in NSDUH 2018 is statistically significant.

[b]According to the Appendix A in NSDUH 2018 (35), Glossary, "Although the specific pain relievers listed above are classified as opioids, use or misuse of any other pain reliever could include prescription pain relievers that are not opioids. For misuse in the past year or past month, estimates could include small numbers of respondents whose only misuse involved other drugs that are not opioids."

Moving forward, there are two directions to further improve the pipeline, inclusion of targeted cohort into training data and experimenting with additional classification algorithms/architectures. For example, incorporating multiple features in one system, similar to the M3 system,[31] might further improve the performance. We chose our architecture based on model simplicity and development efficiency. We note that, though potentially complex and time-consuming, it is possible to design and train a model that learns from all the user's attributes simultaneously and performs well, in contrast to our architecture that learns these information through a transformed knowledge—the predicted scores. We leave this investigation to future work.

Potential pipeline utility

Given that our pipeline performs well across domains and shows promising results on the external task (eg, inferring gender proportions), we believe that this pipeline is well-suited for application on medical/health tasks harnessing Twitter data. This pipeline can be used to infer the gender proportions in targeted cohorts and potentially help investigate the gender disparities in health topics of interest. For example, social media has been shown to be a potentially excellent resource for conducting large-scale mental health surveillance,[19,53,54] and our methods can be used to derive gender-specific insights from such surveillance tasks. Tasks commonly performed using social media data, such as sentiment analyses regarding targeted topics, may also benefit from the gender-specific characterizations enabled by our system.[55,56] Combined with other recently developed methods, such as geolocation-based characterization of social media chatter,[14,50] our methods can provide very unique insights over a given population of social media users.

Toxicovigilance

Our post-classification analyses of the PM cohort illustrated the utility of automatic gender classification on social media data. The closeness of the gender proportions of tranquilizer and stimulant misusers from Twitter and those from NSDUH 2018 validates the effectiveness applying social media mining for Toxicovigilance.[10,57,58] The inferred gender proportion of pain reliever users, though different from NSDUH 2018, is almost identical to that of the overdose EDV according to the NEDS. This association between self-reports of drug use on Twitter and overdose EDV rates is consistent with our past research,[14] in which we identified significant associations between opioid misuse reports on Twitter and overdose deaths over specific geolocations (eg, counties and sub-states). Social media provides the opportunity to combine multiple types of information, including past tweets, social connections, and geolocation. All the information combined can provide geolocation-, gender- and time-specific trends to extract insights, for example for gender inequalities in medical treatment regarding substance use disorder.[59–63] It potentially could also test hypotheses such as the association between mental health issues and PM misuse.[64,65] The development of sophisticated models for social media mining may even provide broad insights about how nonmedical users of pain relievers become victims of overdose over time, and may even serve as an early warning system.[57,58,66–68] Furthermore, the surveillance can be done close to real time—a great improvement over the turnaround time for curating overdose statistics and conducting the NSDUH, which may make timely public health intervention possible.[69] For example, the system can provide the trends and statistics to the local health department and hospitals for better preparation for PM misuse prevention and treatment, and highlight cohorts at higher risk.[57,70] Note that we do not envision that social media data analytics can replace the traditional resources, but we know from the current state of research that it provides excellent complementary data, and the opportunity to provide information/intervention beyond the traditional health services.

Limitations

Our pipeline may inherit the biases and errors introduced by the data and resources used in the pipeline development, leading to significant limitations. The lack of information related to the biases (eg, race, primary language, or location) limits the performance and our ability to address them. For example, the users in Dataset-1, though having at least one English tweet, may not be representative for US Twitter users (eg, by racial distribution). Our pipeline may inherit this undetected bias. Also, using Demographer[28] might introduce bias toward racial majority. Though its effect on the test performance might be detected, we are not able to remedy such biases when the racial coding is absent. Also biases might be introduced during annotation. For example, Dataset-1 and the test set of Dataset-2 may be biased toward those whose gender identities are public. Therefore, though the evaluation provides a measure of the pipeline's performance against human interpretation, it may not be accurate on users whose genders are difficult to identify.

Besides, merging the two individually labeled datasets when constructing Dataset-1, though essential for obtaining acceptable training power and generalizability, could also affect annotation quality by introducing inconsistency. Though the annotation methods adopted in Liu and Ruths[25] and Burger et al[23] both fall within our definition (gender identified on social media), we caution that these methods are different and are not perfect. For example, some users might use different gender identities on different platforms.

Crucially, limited by the annotations, our methods are only applicable to populations with binary gender identities. While this cov-

ers the majority of the population, our methods do not work for the non-binary gender minorities—a community that has been shown to be particularly vulnerable from a public health perspective.[71–74] Despite this limitation, our proposed system not only serves as an important stepping stone for future work by establishing a strong performance over the simplistic binary classification, but already allows us to investigate the inequalities that women experience in medical treatment (eg, for substance use disorder).[59–63] Including non-binary population in our model would require collecting data from this population using coding schemes tailored for the differences within the population. Obtaining comprehensive demographic information could also help extending our methods to include non-binary users. We currently are in the early stage of exploring how to best address these issues.

There are also significant limitations associated with the analysis of nonmedical PM users. First, not all people living in the United States use English primarily over social media. Limited by our infrastructure, we currently are unable to capture Twitter users who use languages other than English, but extending to other languages, specifically Spanish, is a planned future direction.[75–77] Second, Twitter users might choose not to include geo information in tweets, which makes geo-locating impossible. For example, Dredze et al[50] estimated that only less than 25% of the public tweets could be geo-located by their system. We caution that, because of this low proportion, it is not clear if the tweets geo-located in the United States can well represent the US tweets. For Dataset-2, we found that roughly 40% of the users' misuse tweets could be geo-located while about 84% of them were located in the United States, and the gender proportions inferred using Dataset-2 and those using Dataset-2-US are very similar. Though this suggests that they might represent similar populations, they may still not be representative of all US Twitter users. Third, the detection of misuse tweets is based on a classification pipeline, so the inference is also limited by this NLP pipeline's performance.[49] Fourth, the data are limited to Twitter users that are accessible via the Twitter API, and should not be considered as a random sample of US population.

Ethics

Though we limit this work to observational research on publicly available data and adhere to Twitter API's use terms, there is still concern over Twitter users' protection and their perceptions.[78–81] To avoid potential harms to the users, we only study and report on the aggregated data; no user's data will be released. We also will make the NLP pipeline publicly available (without the data) to ensure reproducibility, transparency to researchers and Twitter users, and to support community-driven development. Only the scripts for gender detection pipeline and our best performing pipeline will be made available with this manuscript.

CONCLUSIONS

As social media-based health research focus is moving from population-level to cohort-level studies, incorporating user demographic information is becoming more important. In this work, we developed a gender detection pipeline and evaluated its performance on a general dataset and a domain-specific dataset. Our proposed pipeline shows high accuracy even when applied on a health-specific dataset. We further showed that the pipeline can be used to infer the nonmedical PM users' gender distributions, which is consistent with the statistical data reported by NSDUH 2018 (stimulants and tranquilizers) and by NEDS (overdose EDV due to Opioids). With the much-needed growing attention on explicitly incorporating demographic information, such as gender and race/ethnicity, in research, it is crucial to be able to conduct aggregated gender-specific analyses of health-related social media data. Our pipeline is readily usable by social media researchers who need to infer users' demographics from their data. We note that, besides gender, other demographic information, such as race or age are also important for research, and developing pipelines for these user characterization tasks and evaluating them on domain-specific datasets are part of our planned future work.

FUNDING

Research reported in this publication was supported by the National Institute on Drug Abuse (NIDA) of the National Institutes of Health (NIH) under award number R01DA046619. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

AUTHOR CONTRIBUTIONS

YY conducted and directed the machine learning experiments, evaluations and data analyses, with assistance from MAA and AS. AS provided supervision for full study. JSL and JP provided toxicology domain expertise for interpreting the results. YY drafted the manuscript and all authors contributed to the final manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGEMENTS

The authors thank the support from the National Institute of Health and National Institute of Drug Abuse.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

The data underlying this article cannot be shared publicly due to Twitter API's use terms and privacy concern. The data will be shared on reasonable request to the corresponding author.

REFERENCES

1. Grajales FJ III, Sheps S, Ho K, Novak-Lauscher H, Eysenbach G. Social media: a review and tutorial of applications in medicine and health care. *J Med Internet Res* 2014; 16 (2): e13.
2. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. *J Med Internet Res* 2013; 15 (4): e85.
3. Fox S. The social life of health information. Pew Research Center. Updated January 15, 2014. https://www.pewresearch.org/fact-tank/2014/01/15/the-social-life-of-health-information/. Published 2014. Accessed February 18, 2021.
4. Yang Y-C, Al-Garadi MA, Bremer W, Zhu JM, Grande D, Sarker A. Developing an Automatic System for Classifying Chatter About Health Serv-

ices on Twitter: Case Study for Medicaid. *J Med Internet Res* 2021; 23 (5): e26616.
5. Glover M, Khalilzadeh O, Choy G, Prabhakar AM, Pandharipande PV, Gazelle GS. Hospital evaluations by social media: a comparative analysis of facebook ratings among performance outliers. *J Gen Intern Med* 2015; 30 (10): 1440–6.
6. Campbell L, Li Y. Are Facebook user ratings associated with hospital cost, quality and patient satisfaction? A cross-sectional analysis of hospitals in New York State. *BMJ Qual Saf* 2018; 27 (2): 119–29.
7. Hefele JG, Li Y, Campbell L, Barooah A, Wang J. Nursing home Facebook reviews: who has them, and how do they relate to other measures of quality and experience? *BMJ Qual Saf* 2018; 27 (2): 130–9.
8. Ranard BL, Werner RM, Antanavicius T, *et al*. Yelp reviews of hospital care can supplement and inform traditional surveys of the patient experience of care. *Health Affairs* 2016; 35 (4): 697–705.
9. Broniatowski DA, Paul MJ, Dredze M. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. *PLoS One* 2013; 8 (12): e83672.
10. Sarker A, O'Connor K, Ginn R, *et al*. Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from Twitter. *Drug Saf* 2016; 39 (3): 231–40.
11. O'Connor K, Pimpalkhute P, Nikfarjam A, Ginn R, Smith KL, Gonzalez G. Pharmacovigilance on twitter? Mining tweets for adverse drug reactions. In: *AMIA annual symposium proceedings*, Vol. 2014. American Medical Informatics Association; 2014: 924.
12. Mowery J. Twitter influenza surveillance: quantifying seasonal misdiagnosis patterns and their impact on surveillance estimates. *Online J Public Health Inform* 2016; 8 (3): e198.
13. Sarker A, Chandrashekar P, Magge A, Cai H, Klein A, Gonzalez G. Discovering cohorts of pregnant women from social media for safety surveillance and analysis. *J Med Internet Res* 2017; 19 (10): e361.
14. Sarker A, Gonzalez-Hernandez G, Ruan Y, Perrone J. Machine learning and natural language processing for geolocation-centric monitoring and characterization of opioid-related social media chatter. *JAMA Netw Open* 2019; 2 (11): e1914672.
15. Al-Garadi MA, Yang Y-C, Lakamana S, *et al*. Automatic breast cancer survivor detection from social media for studying latent factors affecting treatment success. In: Michalowski M, Moskovitch R, eds. Artificial Intelligence in Medicine. AIME 2020. Lecture Notes in Computer Science. Vol. 12299. Cham: Springer; 2020. 10.1007/978-3-030-59137-3_10
16. Coppersmith G, Leary R, Crutchley P, Fine A. Natural language processing of social media as screening for suicide risk. *Biomed Inf Insights* 2018; 10: 1178222618792860.
17. Mowery DL, Park YA, Bryan C, Conway M. Towards automatically classifying depressive symptoms from Twitter data for population health. In: Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES); The COLING 2016 Organizing Committee; 2016: 182–91.
18. Coppersmith G, Dredze M, Harman C, Hollingshead K. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In: Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality; Association for Computational Linguistics; 2015: 1–10.
19. Amir S, Dredze M, Ayers JW. Mental health surveillance over social media with digital cohorts. In: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology; Association for Computational Linguistics; 2019: 114–20.
20. Cesare N, Grant C, Nguyen Q, Lee H, Nsoesie EO. How well can machine learning predict demographics of social media users? *arXiv Preprint arXiv:170201807*. 2017.
21. Cesare N, Grant C, Hawkins JB, Brownstein JS, Nsoesie EO. Demographics in social media data for public health research: does it matter? Bloomberg Data for Good Exchange Conference. New York; 2017.
22. Inclusion of women and minorities as participants in research involving human subjects. https://grants.nih.gov/policy/inclusion/women-and-minorities.htm. Accessed August 25, 2020.
23. Burger JD, Henderson J, Kim G, Zarrella G. Discriminating gender on Twitter. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics; 2011: 1301–9.
24. Alowibdi JS, Buy UA, Yu P. Language independent gender classification on Twitter. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '13). New York, NY, USA: Association for Computing Machinery; 739–43. DOI:https://doi.org/10.1145/2492517.2492632
25. Liu W, Ruths D. What's in a name? using first names as features for gender inference in twitter. In: 2013 AAAI Spring Symposium Series. Association for the Advancement of Artificial Intelligence; 2013.
26. Sap M, Park G, Eichstaedt J, *et al*. Developing age and gender predictive lexica over social media. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); Association for Computational Linguistics; 2014: 1146–51.
27. Merler M, Cao L, Smith JR. You are what you tweet. . . pic! gender prediction based on semantic analysis of social media images. In: 2015 IEEE International Conference on Multimedia and Expo (ICME); IEEE; 2015: 1–6.
28. Knowles R, Carroll J, Dredze M. Demographer: Extremely Simple Name Demographics. In: Proceedings of the First Workshop on NLP and Computational Social Science; Association for Computational Linguistics; 2016: 108–13.
29. Bsir B, Zrigui M. Bidirectional LSTM for author gender identification. In: Nguyen N., Pimenidis E., Khan Z., Trawiński B. (eds) Computational Collective Intelligence. ICCCI 2018. Lecture Notes in Computer Science, vol 11055. Springer, Cham. 10.1007/978-3-319-98443-8_36
30. Vicente M, Batista F, Carvalho JP. Gender detection of Twitter users based on multiple information sources In: Kóczy L, Medina-Moreno J, Ramírez-Poussa E, eds. Interactions Between Computational Intelligence and Mathematics Part 2. Studies in Computational Intelligence. Vol 794. Cham: Springer. 10.1007/978-3-030-01632-6_3
31. Wang Z, Hale S, Adelani DI, *et al*. Demographic inference and representative population estimates from multilingual social media data. In: The World Wide Web Conference (WWW '19). Association for Computing Machinery; New York, NY, USA; 2056–2067. DOI:https://doi.org/10.1145/3308558.3313684
32. Zhang C, Abdul-Mageed M. BERT-based arabic social media Author-Profiling. In: Mehta P, Rosso P, Majumder P, Mitra M, eds. Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings. CEUR-WS.org; December 12–15, 2019; Kolkata, India.
33. Volkova S, Wilson T, Yarowsky D. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2013: 1815–27.
34. Huang X, Smith MC, Paul MJ, *et al*. Examining patterns of influenza vaccination in social media. In: AAAI Workshops. Association for the Advancement of Artificial Intelligence; 2017.
35. Substance Abuse and Mental Health Services Administration. Results from the 2018 National Survey on Drug Use and Health: Detailed tables. Rockville, MD: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration; 2019. https://www.samhsa.gov/data/.
36. Centers for Disease Controland Prevention. 2019 AnnualSurveillance Report of Drug-RelatedRisks and Outcomes — UnitedStates Surveillance Special Report.Centers for Disease Control and Prevention, U.S. Departmentof Health and Human Services.Published November 1, 2019.
37. Amazon Mechanical T. https://www.mturk.com/. Accessed November 6, 2020.
38. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol* 2011; 2 (3): 1–27.
39. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv Large Margin Classifiers* 1999; 10 (3): 61–74.

40. Ho TK. Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition. vol. 1. IEEE; 1995: 278–82.
41. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997; 9 (8): 1735–80.
42. Schuster M, Paliwal KK. Bidirectional recurrent neural networks. *IEEE Trans Signal Process* 1997; 45 (11): 2673–81.
43. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Vol. 1 (Long and Short Papers); Association for Computational Linguistics; 2019: 4171–86.
44. Liu Y, Ott M, Goyal N, *et al.* Roberta: a robustly optimized bert pretraining approach. *arXiv preprint arXiv:190711692*. 2019.
45. Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. In: Advances in Neural Information Processing Systems; Curran Associates, Inc; 2017.
46. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. *OpenAI Blog* 2019; 1 (8): 9.
47. Conneau A, Lample G. Cross-lingual language model pretraining. In: Advances in Neural Information Processing Systems; Curran Associates, Inc; 2019.
48. O'Connor K, Sarker A, Perrone J, Hernandez GG. Promoting reproducible research for characterizing nonmedical use of medications through data annotation: description of a Twitter corpus and guidelines. *J Med Internet Res* 2020; 22 (2): e15861.
49. Al-Garadi MA, Yang Y-C, Cai H, *et al.* Text classification models for the automatic detection of nonmedical prescription medication use from social media. *BMC Med Inf Decis Mak* 2021; 21 (1): 27.
50. Dredze M, Paul MJ, Bergsma S, Tran H. , Carmen: A twitter geolocation system with applications to public health. In: AAAI workshop on expanding the boundaries of health informatics using AI (HIAI). Vol. 23; Carmen: A twitter geolocation system with applications to public health; Citeseer; 2013: 45.
51. Abuse NIoD. Research report series. *Prescription Drugs—Abuse and Addiction* 2001.
52. Heil B, Piskorski M. New Twitter research: men follow men and nobody tweets. *Harv Bus Rev* 2009; 1: 2009.
53. Birnbaum ML, Ernala SK, Rizvi AF, De Choudhury M, Kane JM. A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals. *J Med Internet Res* 2017; 19 (8): e289.
54. Reece AG, Reagan AJ, Lix KL, Dodds PS, Danforth CM, Langer EJ. Forecasting the onset and course of mental illness with Twitter data. *Sci Rep* 2017; 7 (1): 1–11.
55. Zunic A, Corcoran P, Spasic I. Sentiment analysis in health and well-being: systematic review. *JMIR Med Inform* 2020; 8 (1): e16023.
56. Gohil S, Vuik S, Darzi A. Sentiment analysis of health care Tweets: review of the methods used. *JMIR Public Health Surveill* 2018; 4 (2): e43.
57. Chary M, Genes N, McKenzie A, Manini AF. Leveraging social networks for toxicovigilance. *J Med Toxicol* 2013; 9 (2): 184–91.
58. Sarker A, DeRoos A, Perrone J. Mining social media for prescription medication abuse monitoring: a review and proposal for a data-centric framework. *J Am Med Inf Assoc* 2020; 27 (2): 315–29.
59. McHugh RK, Votaw VR, Sugarman DE, Greenfield SF. Sex and gender differences in substance use disorders. *Clin Psychol Rev* 2018; 66: 12–23.
60. Manuel JI, Lee J. Gender differences in discharge dispositions of emergency department visits involving drug misuse and abuse—2004-2011. *Subst Abuse Treat Prev Policy* 2017; 12 (1): 1–12.
61. Ryoo H-J, Choo EK. Gender differences in emergency department visits and detox referrals for illicit and nonmedical use of opioids. *WestJEM* 2016; 17 (3): 295–301.
62. Beaudoin FL, Baird J, Liu T, Merchant RC. Sex differences in substance use among adult emergency department patients: prevalence, severity, and need for intervention. *Acad Emerg Med* 2015; 22 (11): 1307–15.
63. Choo EK, Douriez C, Green T. Gender and prescription opioid misuse in the emergency department. *Acad Emerg Med* 2014; 21 (12): 1493–8.
64. Hawkins EH. A tale of two systems: co-occurring mental health and substance abuse disorders treatment for adolescents. *Ann Rev Psychol* 2009; 60: 197–227.
65. Unger JB, Kipke MD, Simon TR, Montgomery SB, Johnson CJ. Homeless youths and young adults in Los Angeles: prevalence of mental health problems and the relationship between mental health and substance abuse disorders. *Am J Commun Psychol* 1997; 25 (3): 371–94.
66. Kenne DR, Hamilton K, Birmingham L, Oglesby WH, Fischbein RL, Delahanty DL. Perceptions of harm and reasons for misuse of prescription opioid drugs and reasons for not seeking treatment for physical or emotional pain among a sample of college students. *Subst Use Misuse* 2017; 52 (1): 92–9.
67. Boys A, Marsden J, Strang J. Understanding reasons for drug use amongst young people: a functional perspective. *Health Educ Res* 2001; 16 (4): 457–69.
68. Stewart SH, Karp J, Pihl RO, Peterson RA. Anxiety sensitivity and self-reported reasons for drug use. *J Subst Abuse* 1997; 9: 223–40.
69. Cao B, Gupta S, Wang J, *et al.* Social media interventions to promote HIV testing, linkage, adherence, and retention: systematic review and meta-analysis. *J Med Internet Res* 2017; 19 (11): e394.
70. Sloboda Z. Changing patterns of "drug abuse" in the United States: connecting findings from macro-and microepidemiologic studies. *Subst Use Misuse* 2002; 37 (8–10): 1229–51.
71. Meerwijk EL, Sevelius JM. Transgender population size in the United States: a meta-regression of population-based probability samples. *Am J Public Health* 2017; 107 (2): e1–e8.
72. Mayer KH, Bradford JB, Makadon HJ, Stall R, Goldhammer H, Landers S. Sexual and gender minority health: what we know and what needs to be done. *Am J Public Health* 2008; 98 (6): 989–95.
73. Streed CG, McCarthy EP, Haas JS. Association between gender minority status and self-reported physical and mental health in the United States. *JAMA Intern Med* 2017; 177 (8): 1210–2.
74. Reisner SL, Greytak EA, Parsons JT, Ybarra ML. Gender minority social stress in adolescence: disparities in adolescent bullying and substance use by gender identity. *J Sex Res* 2015; 52 (3): 243–56.
75. Soares F, Villegas M, Gonzalez-Agirre A, Krallinger M, Armengol-Estapé J. Medical word embeddings for Spanish: Development and evaluation. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop; Association for Computational Linguistics; 2019: 124–33.
76. Segura-Bedmar I, Martínez P, Revert R, Moreno-Schneider J. Exploring Spanish health social media for detecting drug effects. *BMC Med Inform Decis Mak* 2015; 15 (Suppl 6): 1–9. 10.1186/1472-6947-15-S2-S6.
77. Cook BL, Progovac AM, Chen P, Mullin B, Hou S, Baca-Garcia E. Novel use of natural language processing (NLP) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in Madrid. *Comput Math Methods Med* 2016; 2016: 1–8.
78. Williams ML, Burnap P, Sloan L. Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology* 2017; 51 (6): 1149–68.
79. Mello MM, Wang CJ. Ethics and governance for digital disease surveillance. *Science* 2020; 368 (6494): 951–4.
80. Klingwort J, Schnell R. Critical Limitations of Digital Epidemiology. *Surv Res Methods* 14 (2): 95–101. 10.18148/srm/2020.v14i2.7726
81. Morgan HM. Research note: surveillance in contemporary health and social care: friend or foe? *Surveill Soc* 2014; 12 (4): 594–6.