Short Communication

WOMEN'S HEALTH

Women's Health
Volume 16: 1–6
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1745506520949392
journals.sagepub.com/home/whe
SAGE

Studying social media language changes associated with pregnancy status, trimester, and parity from medical records

Sharath Chandra Guntuku[1,2,3], Jessica S Gaulton[1,2], Emily K Seltzer[1,4], David A Asch[1,4], Sindhu K Srinivas[5], Lyle H Ungar[1,3,6], Christina Mancheno[1], Elissa V Klinger[1,4] and Raina M Merchant[1,2,4]

Abstract

We sought to evaluate whether there was variability in language used on social media across different time points of pregnancy (before, during, and after pregnancy, as well as by trimester and parity). Consenting patients shared access to their individual Facebook posts and electronic medical records. Random forest models trained on Facebook posts could differentiate first trimester of pregnancy from 3 months before pregnancy (F1 score = .63) and from a random 3-month time period (F1 score = .64). Posts during pregnancy were more likely to include themes about family (β = .22), food craving (β = .14), and date/times (β = .13), while posts 3 months prior to pregnancy included themes about social life (β = .30), sleep (β = .31), and curse words (β = .27), and 3 months post-pregnancy included themes of gratitude (β = .17), health appointments (β = .21), and religiosity (β = .18). Users who were pregnant for the first time were more likely to post about lack of sleep (β = .15), activities of daily living (β = .09), and communication (β = .08) compared with those who were pregnant after having a child who posted about others' birthdays (β = .16) and life events (.12). A better understanding about social media timelines can provide insight into lifestyle choices that are specific to pregnancy.

Keywords

language, pregnancy, social media, trimester, Facebook, machine learning

Date received: 27 November 2019; revised: 26 March 2020; accepted: 20 July 2020

Introduction

With the ubiquitous nature of smart devices and Internet in many regions of the world, users are increasingly using social media to share information about their lifestyle. Many Internet users spend upward of 10% of time of each day using social media platforms like Facebook, Twitter, Instagram, and Snapchat. Approximately, 15%–25% of this content is health-related.[1] Social media can provide clinicians with unique insight into the day-to-day experiences and behaviors of individuals.[2] Prior work has shown that pregnancy is a time when women rely on information acquired from social media sites to inform their health and lifestyle choices.[3,4] Online platforms are also heavily used for social support before, during, and after pregnancy[5,6]

[1]Penn Medicine Center for Digital Health, University of Pennsylvania, Philadelphia, PA, USA
[2]Department of Emergency Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[3]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA
[4]Penn Medicine Center for Health Care Innovation, University of Pennsylvania, Philadelphia, PA, USA
[5]Department of Obstetrics and Gynecology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[6]Positive Psychology Center, University of Pennsylvania, Philadelphia, PA, USA

Corresponding author:
Sharath Chandra Guntuku, Department of Computer and Information Science, University of Pennsylvania, 3330 Walnut St, Philadelphia, PA 19104, USA.
Email: sharathg@seas.upenn.edu

and could be a powerful tool to gain insight into lifestyle choices that were previously unobservable.

There is evidence that social media language is indicative of individuals' psychological stress,[7,8] loneliness,[9] depression,[10] suicide risk,[11] hospital utilization,[12] and medical conditions in general.[13] Prior works on studying pregnancy using social media focused on predicting the risk of postpartum depression[14] and identifying women who have announced their pregnancy on Twitter to detect large numbers of cohorts.[15] Instagram has been used to study the attitudes of pregnant women toward pregnancy and how their posts reflect it.[16,17] To our knowledge, no other studies have investigated variability in social media language across the time period of pregnancy by individuals as noted by their medical records.

As a first step to understanding whether data from social media can provide insights about lifestyle in pregnancy, we aimed to use machine learning to evaluate how language differed across pregnancy status, trimester, and parity to better understand the experiences of pregnant women as shared online. We also evaluated whether a machine learning model can distinguish language of the first trimester compared with a random time period and 3 months prior conception.

Materials and methods

This study was approved by the University of Pennsylvania Institutional Review Board. This was a retrospective analysis of social media and electronic medical records (EMR) data of consenting patients with a prior history of delivery.

Data

Data were collected from an ongoing study (March 2014 to present) of merged social media and EMR from consenting patients for research.[18,19] Using a convenience sample framework, adult patients receiving care in outpatient (i.e. internal medicine clinics), acute care (i.e. emergency department), and inpatient (i.e. postpartum suite) settings of an urban academic hospital system were approached for study participation. Historical data that were posted on social media from the time they signed up on Facebook was extracted and stored securely. No data were accessed from the Facebook pages of study participants' friends or from posts on the study participants' page made by anyone other than the participant. At the time of enrollment, participants also completed a brief survey about demographics, social media use, and health status.

Of patients who endorsed using social media (n = 16,507) and indicated that they were willing to participate in research (n = 6872), 5885 (86%) consented to share access to their social media (e.g. Facebook, Twitter, and/or Instagram) data and EMR data.

Study population

From the study cohort, we identified female patients with a history of a delivery in our health system EMR. Demographic data for each participant were extracted from self-reported survey (e.g. age, gender, and race). We then identified delivery dates for these participants which coincided with the time when they also had status updates on Facebook before, during, and after pregnancy.

Identifying differences in the content of social media posts by pregnancy status and parity

To identify themes in language during the defined time periods, we extracted the relative frequency of single words across all Facebook status posts of all users.[20] We removed all words used by less than 1% of users. We extracted topics (clusters of co-occurring words) from all messages posted by the users in the cohort using unsupervised latent Dirichlet allocation (LDA). The LDA generative model[21] assumes that posts contain a combination of topics and that topics are a distribution of words. As the words in a post are known, topics, which are latent variables, can be estimated through Gibbs sampling. We used the Mallet implementation, adjusting one parameter ($\alpha = 5$) to favor fewer topics per post. We used the default English stop-words provided in the Mallet package. In our study, 200 topics were generated using tweets across all users in the dataset. If removing infrequent words prior to LDA would result in less than 400 words for a user, that user was still included in the analysis.

We used logistic regression to distinguish the different features associated with (a) pre-conception (3 months prior conception), during pregnancy, and postpartum (3 months after childbirth) compared with a random time period in the patient's social media archive and (b) with parity (characterized binarily as one live birth or more than one live birth as determined by a documentation of a prior delivery in our health system). We measured the effect size using the regression coefficients (β). For identifying themes from topics, researchers looked at 20 messages each with the highest topic prevalence. For example, a topic including the words ("hungry," "cook," "starving," "food") was labeled as a hungry topic, ("sleepy," "tired," "bed," "sleep," "ugh," "gettin") was identified as a sleep topic. We used Benjamin–Hochberg p-correction and $p < .01$ for indicating meaningful associations.

We did not have enough data to obtain topics with significant differences to do a per trimester analysis on themes. The average number of words across all participants was 100 for first trimester, 90 for second and 120 for third.

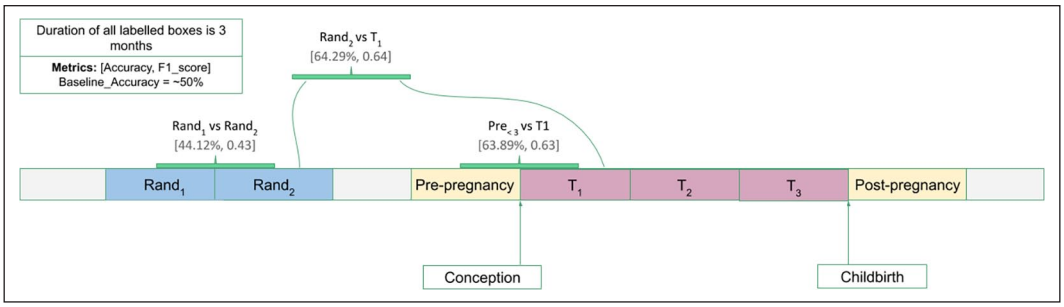

Figure 1. Performance of Random Forest classifier trained on LDA topics extracted from Facebook posts at differentiating phases of pregnancy. T_n denotes different trimesters (n=1, 2, 3). Pre- and Post-pregnancy denote 3 months prior conception and post childbirth respectively. $Rand_n$ denotes a random 3 month window not during pregnancy.

Predicting pregnancy trimester and parity using language features

We took a case-crossover approach[12] to evaluate if machine learning models would distinguish time periods within individuals relevant to pregnancy—primarily the first trimester where developing interventions to improve early prenatal care can substantially improve pregnancy outcomes. We used the LDA topics as features to train a Random Forest classifier to distinguish linguistic cues associated with first-trimester pregnancy compared with a random time period and 3 months before conception (see Figure 1). We also tested with the bag of words features, but found them to be inferior in performance compared with the LDA topics. To identify differences by pregnancy status, we compared language features during first trimester with those during a random 3-month time window in the patient's social media archive and with pre-pregnancy (3 months prior to estimated conception date).

All models were trained on 80% of the participants and tested on the held-out sample of 20%. Thus, the predictive model was created without any outcome information outside of the training data, making the test data an out-of-sample evaluation (i.e. participants in the training set were not a part of the test set). Random Forest classifier from the scikit-learn module in Python 3.4 was used to fit on the training set using five-fold cross-validation to tune the parameters. The trained model was then applied on the test set to report predictive performance. We measured the performance using accuracy and F1 scores due to class balance. Since this was a case-crossover study design, every patient was acting as their own control, that is, for instance, the language in the random event is considered as control and language in the first trimester is considered a part of the case group.

Results

Study participants

Of 654 female patients, 72% (471) met study inclusion criteria. These participants, were primarily young (median age: 25 ± 6 years) and African American 85% (400; Table 1). From these patients, we computationally analyzed 505,114 posts.

Table 1. Characteristics of the study cohort.

Descriptive statistics	
Demographic characteristic	Number of participants
Female	471
Race	
African American	400 (85%)
White	58 (12%)
Other	14 (3%)
Age range (years)	19–42
Median age (years)	24.6

Content of social media posts by pregnancy status and parity

Pregnancy status. Language during pregnancy differed significantly from language prior to pregnancy and post-pregnancy (Table 2). During pregnancy, participants posted more medical terms—"pain," "hospital," "doctor," —(β=.107), about having food cravings—"hungry," "cook," "starving," "food"—(β=.145), about family—"care," "father," "parents," "babies"—(β=.222), and had more posts about dates and times—"last," "year," "first," "month," "weeks," "time"—(β=.134). Pre-pregnancy, participants were mostly posting about being relaxed—"bored," "chill," "house," "hmu"—(β=.34), about sleep—"bed," "dreams," "goodnite," "rest," "sleepy"—(β=.31), having a social life—"music," "party," "dj," "2nite," "friday"—(β=.306), and are more likely to use curse words (β=.277). Post-pregnancy, participants used language indicative of gratitude—"thanks," "amazing," "blessed," "truly"—(β=.178), religiousity—"god," "jesus," "lord," "pray," "thank," "dear"—(β=.183), and health appointments—"doctors," "appointment," "waiting"—(β=.218), and parent-focused holidays—"mother's," "day," "father's," "day," "special"—(β=.174).

Table 2. Content of social media posts by pregnancy status and parity. All topics are significant at $p < .01$ and Benjamin–Hochberg corrected.

	Theme	Top words in the topic	Regression coefficient (β)
Pregnancy status			
Pre-pregnancy	*relaxing*	bored, chill, house, chillin, hmu	.340
	sleep	bed, dreams, goodnite, rest, sleepy	.310
	social life	music, party, dj, 2nite, friday	.306
	curse words		.277
During pregnancy	*family*	care, father, parents, babies, mommy	.222
	food	hungry, food, cook, breakfast, starving	.145
	date/time	last, year, first, month, weeks, times	.134
	medical	hospital, pain, test, doctor, nurse	.107
Post-pregnancy	*health appointments*	doctors, appointment, waiting	.218
	religious	god, jesus, lord, pray, thank, dear	.183
	gratitude	thanks, amazing, blessed, truly	.178
	parent-focused holidays		.174
Parity			
Parity = 1	lack of sleep	sleepy, tired, bed, sleep, ugh, gettin	.152
	ADLs	dressed, gettin ready, early, dress, shower, bouta	.091
	communication	call, text, somebody, hmu, txt, phone	.089
Parity > 1	birthday (others)	party, bday, month, next, 1st, date	.161
	birthday (self)	thanks, wishes, everyone, thanx, appreciate, special	.124
	life events	wedding, congrats, proud, graduation, mom, amazing, pictures, awesome	.123

ADL: activities of daily living.

Variability in the language on social media with parity status. Language also varied by parity (one vs multiple children). Significant thematic differences are shown in Table 2. Participants who were pregnant for the first time talk a lot more about lack of sleep—"sleepy," "tired," "bed," "sleep," "ugh," "gettin"—(β = .152), about communicating—"call," "text," "somebody," "hmu," "txt," "phone"—(β = .089), and about activities of daily living (ADL)—"dressed," "gettin ready," "early," "dress," "shower"—(β = .091), whereas participants who were pregnant after having a child posted a lot more about birthdays, both their own—"thanks," "wishes," "everyone," "thanx," "appreciate," "special"—(β = .124), and of others—"party," "bday," "month," "next," "1st"—(β = .161), and life events—"wedding," "congrats," "proud," "graduation," "mom," "pictures"—(β = .123).

Predictive analysis

Random forest models could differentiate posts during the first trimester of pregnancy from those prior to pregnancy (F1 score = .63) and from posts during a random 3-month time period (F1 score = .64). Furthermore, to assess whether the model was identifying linguistic differences associated with only random time points, we identified two random consecutive 3-month periods and tested whether the model could predict the difference in linguistic cues between the two 3-month periods. When distinguishing between two random consecutive 3-month periods, the model performs poorly (F1 score = .44) indicating that the model used to detect language associated with the first trimester is identifying differences which extend beyond simple temporal changes.

Discussion

This study's main finding is that there are distinct themes in social media language before, during, and after pregnancy that also differ by parity.

We validated a machine learning model developed from a large social media database merged with the medical records. Validating the machine learning model is just the first step toward understanding whether there are linguistic clues from social media data that can help clinicians understand the motivations behind behaviors that contribute to lifestyle-related disorders of pregnancy. Due to the constraints of health care system, providers have limited time with their patients in the outpatient setting. These time constraints make it difficult to gain insight into the nonmedical aspects of their patient's lives including their social support, lifestyle choices, cultural and religious background, and personal values.

Posts in context

Understanding the linguistic predictors before, during, and after pregnancy can inform the implementation of various interventions. Although this work is exploratory and limited by the size of available data, reviewing the posts we have found many examples of actionable content. An example of a pre-pregnancy intervention may be related to smoking cessation. All examples are paraphrased for anonymity. Individuals often post about their day-to-day habits, including smoking. This is an example post from someone from our cohort within 3 months of getting pregnant. "Kids are off to school now I am about to smoke . . . ."

Monitoring one's diet during pregnancy is also important. Social media may provide more passive data around a pregnant patient's diet, which can supplement self-reported data. One example from our cohort, "I want some seafood . . . can somebody bring me some??" Seafood consumption during pregnancy is recommended, with the caveat of watching mercury levels. Alcohol consumption while pregnant is discouraged, but it is difficult for providers to know if their patients are consuming alcohol due to the stigma and reliance on self-reporting. We found multiple posts related to alcohol during pregnancy. For example, "I know you can't drink liquor when you are pregnant, but can you drink wine or is wine still consider alcohol?" This highlights the complicated and confusing nature of recommendations women hear during pregnancy. Another potential actionable use of these data is around pain management. Many women post about pain and discomfort during their pregnancy, but may not remember to report this to their provider if the episode has passed by the time of their appointment. One participant posted "is having right upper abdominal pain. . . . I think that's a funny update," weeks before her prenatal appointment.

Post-pregnancy can be a stressful period for many women. By evaluating postpartum social media language, providers may be able to detect early signs of psychological stress. An example from our data highlights how language may provide insight into an individual's mental wellbeing, "I definitely can't take another year like this . . . God does put more on you than you can bear." If a provider knew that their patient was experiencing elevated stress, they could provide a referral to a mental health practitioner or social worker.

Identifying themes in language during pregnancy derived from social media can help medical providers gain insight into their patient's lives and can shape anticipatory guidance during the office visit. There may be symptoms of other comorbidities associated with pregnancy, such as gestational hypertension, obesity, preeclampsia, mental health disorders, infection, and so on, that could be identified earlier using social media data. Early prenatal care (first trimester) can improve pregnancy outcomes, provide better gestational age dating, prevent complications, address pre-existing medical conditions, and allow for modification of lifestyle issues such as smoking, drug, alcohol, and excessive weight gain.[22,23] Delayed prenatal care leads to poorer outcomes for high-risk pregnancies, increased chance of cesarean deliveries, and prematurity.[22]

While our analyses did not focus on specific disease processes related to pregnancy, future studies could investigate whether there are themes in social media language that predict the development of specific disorders. For example, among women with gestational hypertension or depression, there may be distinct keywords or linguistic themes that precede this diagnosis that may help identify symptoms even before they see their physician or mental health provider.

Limitations

We included the participants who had Facebook posts across all phases of pregnancy. There might be several participants who post only during some stages of pregnancy, and enforcing the inclusion criteria skews our sample and findings toward social media super utilizers. To better understand their online behavior, data from other sources (e.g. other social media platforms, smartphones, sensors) could be explored in future studies. Although we identified that social media language changes around pregnancy, the specificity needed to confidently use this as a standalone method of diagnosis cannot be concluded at this time.

Conclusion

Posts on social media can be used to predict pregnancy only slightly better than chance and there are distinct themes that pregnant women post about during different stages of pregnancy. A better understanding about these can provide insight into lifestyle choices that are specific to pregnancy. This could be associated with early identification and prevention of risk and could inform implementation of specific interventions during pregnancy.

Acknowledgements

The authors would like to thank Justine Marks, Molly Casey, and Janice Lau for helping with data collection.

Declaration of conflicting interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: David A Asch is a US government employee.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by the Robert Wood Johnson Foundation Pioneer award (Merchant).

ORCID iD

Sharath Chandra Guntuku https://orcid.org/0000-0002-2929-0035

References

1. Sayakhot P and Carolan-Olah M. Internet use by pregnant women seeking pregnancy-related information: a systematic review. *BMC Pregnancy Childbirth* 2016; 16(1): 65.
2. Abnousi F, Rumsfeld JS and Krumholz HM. Social determinants of health in the digital age. *JAMA* 2019; 321(3): 247.
3. Holtz B, Smock A and Reyes-Gastelum D. Connected motherhood: social support for moms and moms-to-be on Facebook. *Telemed J E Health* 2015; 21(5): 415–421.
4. Harpel T. Pregnant women sharing pregnancy-related information on Facebook: Web-Based Survey Study. *J Med Internet Res* 2018; 20(3): e115.
5. Maloni JA, Przeworski A and Damato EG. Web recruitment and Internet use and preferences reported by women with postpartum depression after pregnancy complications. *Arch Psychiatr Nurs* 2013; 27(2): 90–95.
6. Leaver T and Highfield T. Visualising the ends of identity: pre-birth and post-death on Instagram. *Inf Commun Soc* 2018; 21(1): 30–45.
7. Guntuku SC, Buffone A, Jaidka K, et al. Understanding and measuring psychological stress using social media. In: *Proceedings of the 13th international conference on web and social media (ICWSM 2019)*, Munich, 11–14 June 2019.
8. Jaidka K, Guntuku SC, Lee JH, et al. The rural–urban stress divide: obtaining geographical insights through Twitter. *Comput Human Behav* 2020; 114: 106544.
9. Guntuku SC, Schneider R, Pelullo A, et al. Studying expressions of loneliness in individuals using twitter: an observational study. *BMJ Open* 2019; 9: e030355.
10. Guntuku SC, Preotiuc-Pietro D, Eichstaedt JC, et al. What Twitter profile and posted images reveal about depression and anxiety. In: Proceedings of the international AAAI conference on web and social media, vol. 13, Munich, 11–14 June 2019, pp. 236–246. Palo Alto, CA: AAAI Press.
11. Matero M, Idnani A, Son Y, et al. Suicide risk assessment with multi-level dual-context language and BERT, 2019, https://www.aclweb.org/anthology/W19-3005/
12. Guntuku SC, Schwartz HA, Kashyap A, et al. Variability in language used on social media prior to hospital visits. *Sci Rep* 2020; 10: 11456.
13. Merchant RM, Asch DA, Crutchley P, et al. Evaluating the predictability of medical conditions from social media posts. *PLoS ONE* 2019; 14: e0215476.
14. De Choudhury M, Counts S and Horvitz E. Major life changes and behavioral markers in social media. In: Proceedings of the 2013 conference on computer supported cooperative work—CSCW '13, San Antonio, TX, 23–27 February 2013, pp. 1431–1442. New York: ACM Press.
15. Sarker A, Chandrashekar P, Magge A, et al. Discovering cohorts of pregnant women from social media for safety surveillance and analysis. *J Med Internet Res* 2017; 19: e361.
16. Tiidenberg K and Baym NK. Learn it, buy it, work it: intensive pregnancy on Instagram. *Soc Media + Soc* 2017; 3: 1–13.
17. Eagle RB. "Have you tried ginger?": severe pregnancy sickness and intensive mothering on Instagram. *Fem Media Stud* 2019; 19: 767–769.
18. Padrez KA, Ungar L, Schwartz HA, et al. Linking social media and medical record data: a study of adults presenting to an academic, urban emergency department. *BMJ Qual Saf* 2016; 25(6): 414–423.
19. Smith RJ, Crutchley P, Schwartz HA, et al. Variations in Facebook posting patterns across validated patient health conditions: a prospective cohort study. *J Med Internet Res* 2017; 19(1): e7.
20. Schwartz HA, Eichstaedt JC, Kern ML, et al. Personality, gender, and age in the language of social media: the open-vocabulary approach. *PLoS ONE* 2013; 8(9): e73791.
21. Blei DM, Ng AY and Jordan MI. Latent Dirichlet allocation. *J Mach Learn Res* 2003; 3: 993–1022.
22. Shah JS, Revere FL and Toy EC. Improving rates of early entry prenatal care in an underserved population. *Matern Child Health J* 2018; 22(12): 1738–1742.
23. Cohen AW. Scheduling the first prenatal visit: a missed opportunity. *Am J Obstet Gynecol* 2010; 203(3): 192–193.