
This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.

Barcelona, Veronica; Scharp, Danielle; Moen, Hans; Davoudi, Anahita; Idnay, Betina R.; Cato, Kenrick; Topaz, Maxim

Using Natural Language Processing to Identify Stigmatizing Language in Labor and Birth Clinical Notes

Published in:
Maternal and Child Health Journal

DOI:
[10.1007/s10995-023-03857-4](https://doi.org/10.1007/s10995-023-03857-4)

Published: 01/03/2024

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Barcelona, V., Scharp, D., Moen, H., Davoudi, A., Idnay, B. R., Cato, K., & Topaz, M. (2024). Using Natural Language Processing to Identify Stigmatizing Language in Labor and Birth Clinical Notes. *Maternal and Child Health Journal*, 28(3), 578–586. <https://doi.org/10.1007/s10995-023-03857-4>

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.

Using Natural Language Processing to Identify Stigmatizing Language in Labor and Birth Clinical Notes

Veronica Barcelona¹, Danielle Scharp¹, Hans Moen², Anahita Davoudi³, Betina R. Idnay⁴, Kenrick Cato^{1,5}, Maxim Topaz¹

¹ School of Nursing, Columbia University, 560 West 168th St, Mail Code 6, New York, NY 10032, USA

² Department of Computer Science, Aalto University, Espoo, Finland

³ VNS Health, New York, NY, USA

⁴ Department of Biomedical Informatics, Columbia University, New York, NY, USA

⁵ University of Pennsylvania, Philadelphia, PA, USA

✉ Veronica Barcelona, vb2534@cumc.columbia.edu

Abstract

Introduction: Stigma and bias related to race and other minoritized statuses may underlie disparities in pregnancy and birth outcomes. One emerging method to identify bias is the study of stigmatizing language in the electronic health record. The objective of our study was to develop automated natural language processing (NLP) methods to identify two types of stigmatizing language: marginalizing language and its complement, power/privilege language, accurately and automatically in labor and birth notes.

Methods: We analyzed notes for all birthing people >20 weeks' gestation admitted for labor and birth at two hospitals during 2017. We then employed text preprocessing techniques, specifically using TF-IDF values as inputs, and tested machine learning classification algorithms to identify stigmatizing and power/privilege language in clinical notes. The algorithms assessed included Decision Trees, Random Forest, and Support Vector Machines. Additionally, we applied a feature importance evaluation method (InfoGain) to discern words that are highly correlated with these language categories.

Results: For marginalizing language, Decision Trees yielded the best classification with an F-score of 0.73. For power/privilege language, Support Vector Machines performed optimally, achieving an F-score of 0.91. These results demonstrate the effectiveness of the selected machine learning methods in classifying language categories in clinical notes.

Discussion: We identified well-performing machine learning methods to automatically detect stigmatizing language in clinical notes. To our knowledge, this is the first study to use NLP performance metrics to evaluate the performance of machine learning methods in discerning stigmatizing language. Future studies should delve deeper into refining and evaluating NLP methods, incorporating the latest algorithms rooted in deep learning.

Significance:

What is already known on this subject? Traditional informatics methods include natural language processing, and these methods have been increasingly applied to the study of public health problems using electronic health records.

What this study adds? We identified well-performing machine learning methods to automatically identify stigmatizing language in labor and birth clinical notes. These methods have not been applied to labor and birth clinical notes and have the potential to be a powerful tool in examining perinatal health inequities.

Keywords: bias, natural language processing, electronic health records

Introduction

Disparities in adverse pregnancy and birth outcomes are well-documented in the perinatal health literature (Martin & Osterman, 2018). Though racism and discrimination have been suggested as the primary explanatory factor contributing to these disparities (Braveman et al., 2021), other biases related to minoritized statuses and perinatal outcomes have been less studied (Everett et al., 2022; Malouf et al., 2014; Philipsborn et al., 2021; Togioka et al., 2022). Implicit bias among clinicians has been suggested as a contributor to health disparities, broadly defined (Hall et al., 2015). Hospital-level factors, including clinician bias, may play a significant role in unequal care and poor outcomes in birth settings (Minehart et al., 2021), though efforts to reduce bias in these settings have had limited success (Jindal et al., 2022). Operationalizing antiracism efforts in birth settings remains a problem as there is no consistently agreed upon method to identify and measure bias.

Increasingly, researchers have turned to analyzing clinician documentation in the electronic health record (EHR) as an innovative method for identifying potential clinician bias. Stigmatizing language communicates unintended meanings that can perpetuate socially constructed power dynamics and result in biased care (Shattell, 2009). We define stigmatizing language into two broad categories: marginalizing or negative language, and power/privilege or what others term as “positive” language. For example, marginalizing language may reflect identities that are less socially desirable related to skin color, ability, citizenship status, sexual orientation, gender, body size, wealth, marital status and employment (Everett et al., 2022; Malouf et al., 2014; Philipsborn et al., 2021; Togioka et al., 2022). In contrast, power or privilege language may reflect consciously or unconsciously held beliefs reflecting more desirable identities, perpetuating this bias. Such “positive” language has been previously identified (Park et al., 2021) with language characteristics that signal expressions of approval or other positive feelings toward the individual. These narrative descriptors may reflect adherence (or non-adherence) to cultural values and judgments of characteristics related to people’s reproductive rights and demographic characteristics.

A recent qualitative analysis of electronic health records for birthing people has provided examples of stigmatizing language (Barcelona, Scharp, et al., 2023). For example, questioning the hospitalized person’s credibility is a type of marginalizing language that suggests disbelief in the individual’s words (see Table 1). Quotations may be used to indicate this disbelief, such as when describing the report of a family member who stated that the previous baby was ‘born dead’. Another category commonly seen is labeling the individual as difficult: “patient once again advised about warning signs”. In contrast, power/privilege language indicates that the individual possesses more desirable traits, such as “patient reports nurturing marriage with father of baby who is employed as an investment banker”. The use of

stigmatizing language has the potential to be especially harmful given the politicization of reproduction in the U.S. and the subsequently widely held beliefs of the ideal traits a childbearing person should have (Barcelona, Horton, et al., 2023).

Table 1. Examples of stigmatizing language categories, definitions, and examples from clinical notes.

Marginalizing Language	
Category and definition	Example
<u>Questioning the hospitalized individual’s credibility</u> Indicates disbelief in the person’s words	“Patient stated that she did not know she was pregnant until she was 6 months”
<u>Disapproval</u> Indicates disagreement with the individual’s choices or preferences	“She disclosed symptoms to OB doctor who gave her referral to a psychologist, but she did not follow up because ‘she was not in the mood’”
<u>Stereotyping</u> Ascribing behaviors to race, ethnicity, or culture	“patient states baby will sleep in their bed. [social work] intern discussed though this may be cultural, it is important for baby to have his own bed”
<u>Labeling the individual as ‘difficult’</u> Providing examples to show that the person is acting unreasonably difficult	“Pt is nervous that baby is in the NICU and that she won't bond but has already visited 3 times”
<u>Unilateral decisions</u> <u>Emphasizing clinician authority over patient</u>	“Instructed not to sleep with the baby in bed”
<u>Power/privilege language</u>	
Reporting an individual’s identities that reinforce their place in the hierarchy of preferred reproductive characteristics	“[patient] reports having a nurturing 2-year marriage with [father of baby] who is employed as an investment banker” “Husband is a neurosurgeon” “Patient and spouse plan to hire a nanny”

Correctly identifying stigmatizing language in the EHR is important, as it may perpetuate bias against people seeking clinical care by influencing clinician perception (Beach 2021). For example, one study showed that exposure to marginalizing language in clinical notes was associated with more negative attitudes toward the individual and less aggressive management of their pain (Goddu et al., 2018). In addition, federal law now requires that individuals have access to view their clinical notes, thus negative language in documentation may harm people’s trust in their clinician (Fernández et al., 2021; United States Department of Health and Human Services, 2020). As more Black and Latinx individuals read their own EHR notes containing marginalizing language, this may contribute to further loss of trust towards health care providers (Park et al., 2021; Sun et al., 2022). Though there are no published studies examining the associations between stigmatizing language and obstetric outcomes, this work is underway.

Studies examining stigmatizing language use in clinical notes have commonly employed either traditional qualitative or machine learning methods. Machine learning is commonly defined as a computational approach that can learn and adapt (Alpaydin, 2020). This approach adapts by using algorithms and statistical models to analyze and draw inferences from patterns in data. Natural language processing (NLP) is a type of machine learning that applies computational techniques to the analysis of textual data. NLP can help automate the process of identifying stigmatizing language, it may even help identify stigmatizing language more accurately than human review (Goh et al., 2020). It is important to find the best-functioning NLP methods to identify stigmatizing language use to ensure accuracy and allow for replication and extension of findings.

In a recent scoping review of the literature (manuscript under review) conducted in April 2022, we found that only nine studies have been published on stigmatizing language use in EHR clinical notes. Of these, most (n=5) employed traditional qualitative methods to identify and measure stigmatizing language (Fernández et al., 2021; Hoover et al., 2022; Landau et al., 2022; Martin & Stanford, 2020; Park et al., 2021). The remaining four studies used machine learning, specifically NLP methods, to identify and extract stigmatizing words from clinical notes (Alpert et al., 2019; Beach et al., 2021; Himmelstein et al., 2022; Sun et al., 2022). However, no studies on stigmatizing language in clinical notes were conducted in labor and birth settings, as existing research has focused on internal medicine (Beach et al., 2021; Fernández et al., 2021; Himmelstein et al., 2022; Hoover et al., 2022; Park et al., 2021; Sun et al., 2022), oncology (Alpert et al., 2019), psychiatry (Martin & Stanford, 2020), and pediatric (Landau et al., 2022) settings. These studies did not report findings or present analyses on comparing the accuracy of different NLP classification methods for identifying stigmatizing language use, representing an important gap in the literature.

In our previous qualitative work, we developed two broad categories of taxonomy describing marginalizing language and power/privilege language in a sample of clinical notes from labor and birth settings. This study aimed to use our previously developed dataset of expert-reviewed and labeled clinical notes to explore whether NLP methods can automatically identify stigmatizing language.

Methods

For this study, we used data that were previously manually annotated by experts to provide the training and testing set for NLP methods development. We reviewed EHR data for all birthing people >20 weeks' gestation who were admitted for labor and birth to two urban hospitals in the Northeast during 2017. Relevant data were extracted from the Clinical Data Warehouse, and all narrative notes generated by clinicians during the inpatient stay were cleaned (e.g., irrelevant formatting removed) and prepared for analysis. We then tested several NLP methods to identify marginalizing language and power/privilege language use in clinical notes. This research was conducted in accord with prevailing ethical principles, and this secondary analysis of clinical data is in accordance with the Declaration of Helsinki. We received Institutional Review Board approval (AAAT9870) for this study. The workflow is outlined in Figure 1.

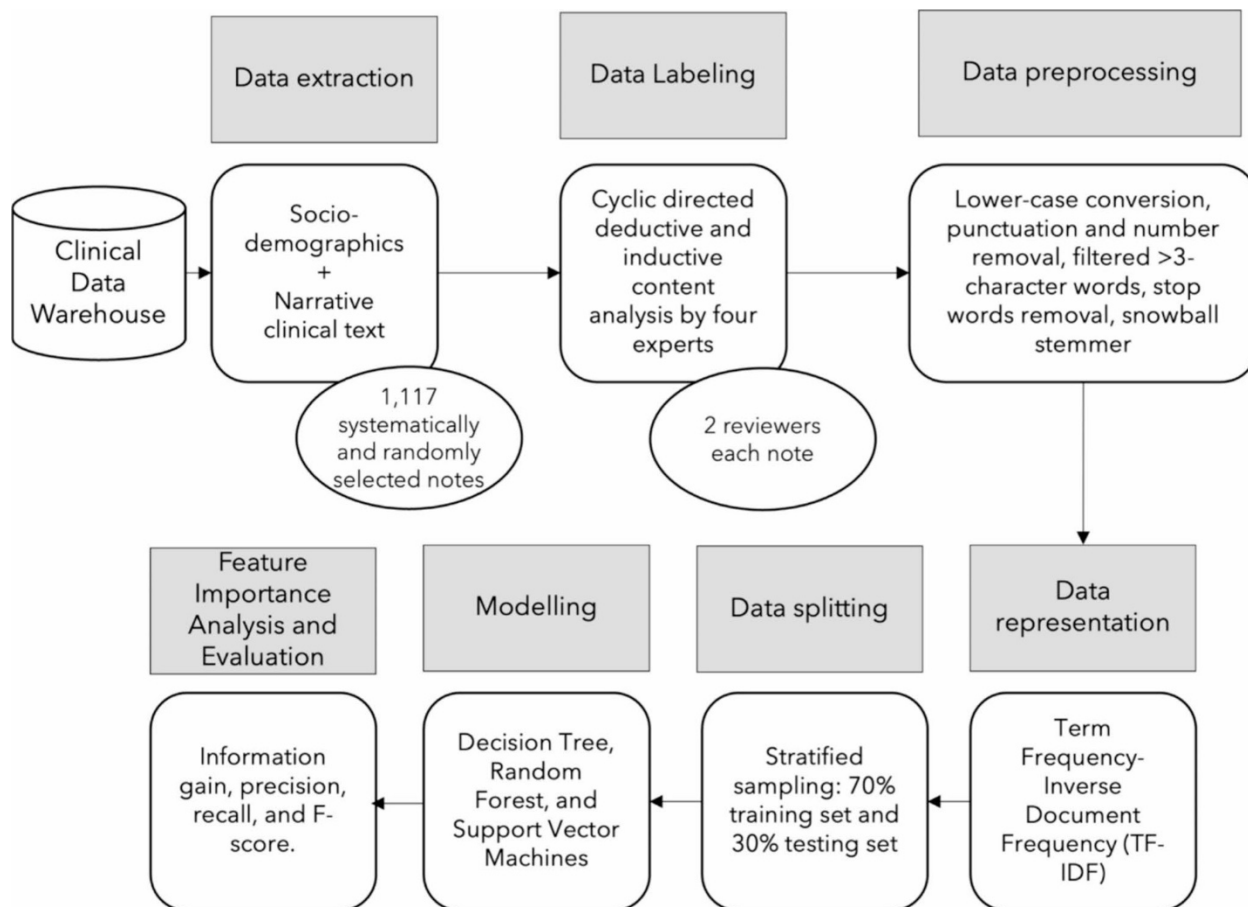


Figure 1. Study workflow.

Manual labeling of clinical notes by experts

We employed qualitative descriptive methodology (Kim et al., 2017; Sandelowski, 2010) to provide a comprehensive summary of the stigmatizing language through theoretical sampling of clinical notes (Coyne, 1997); multiple data sources (note types and clinicians); cyclic directed deductive and inductive content analysis of clinical notes (Hsieh & Shannon, 2005); and condensation of data into thematic representations. Specifically, four members of our research team with expertise in parent/child nursing, data science, and health disparities reviewed a randomly extracted sample of clinical notes stratified by note types. According to our expert panel review, seven note types were likely to have stigmatizing language (i.e., Admission Note, Initial Assessment, Triage Note, Nutrition Assessment, Resident Note, Nursing Note, and Postpartum Note). At least two reviewers reviewed each clinical note, and the review continued until knowledge saturation was achieved after 1,117 notes (i.e., no new categories of stigmatizing language emerged). For each clinical note, reviewers were asked to identify and label any instances of stigmatizing language. The instances of marginalizing language were then divided into the following five categories based on previous literature (Park et al., 2021), including: questioning credibility, disapproval, stereotyping, labeling the person as difficult, and unilateral decisions (Table 1). Several marginalizing language categories were rarely identified in our sample (e.g., only five clinical notes with examples of “unilateral decisions”). Reviewers were also asked to label any new language categories that emerged from the data. We identified power/privilege language that described culturally desirable characteristics, e.g., “husband is an investment banker” or “patient appropriately groomed.”

Interrater agreement was relatively high (Cohen's $K > .8$), and all stigmatizing language cases and categories were agreed upon via research group discussions.

To generate sufficient training and testing samples for NLP methods development, we dichotomized stigmatizing language into two broad categories: marginalizing language and power/privilege language. Overall, 232 (21%) clinical notes had an instance of either marginalizing (n=205, see Table 1 for examples) or power/privilege language (n=37).

NLP methods

We first prepared the clinical notes data using standard text cleaning steps (including lower-casing all words and removing punctuation and numbers) (Tiwary, 2008). Next, the data were transformed into a vector representation using Term Frequency-Inverse Document Frequency (TF-IDF), which builds upon the Bag-of-Words (BoW) representation ((Manning, 2008)). Specifically, TF-IDF weighs each term in a document by its importance relative to the entire corpus. The formula for TF-IDF is given by: $TF - IDF(t, d) = TF(t, d) \times IDF(t)$. Here $TF(t, d)$ is the term frequency of term t in document d , and $IDF(t)$ is the inverse document frequency of term t across the entire corpus. The unit of analysis for our model was the entire note, but the TF-IDF technique was used to convert the text within these notes into a word-level numerical representation. TF-IDF vectorization helps in NLP by converting a collection of documents into numerical vectors, where each dimension represents the importance of a specific word or phrase in the document relative to the entire corpus, allowing for analysis and comparison of the documents based on their relative word frequency and distribution (Tiwary, 2008). These data were split on note level using stratified sampling (stratified by stigmatizing language categories) into 70% training set, and the rest (30%) were kept for testing NLP performance.

Three commonly used machine learning algorithms were tested to identify which performed best at the task of classifying if a note contained marginalizing or power/privilege language. Specifically, we applied Decision Tree (J48), Random Forest, and Support Vector Machines.

- Decision Trees (Quinlan, 2014) can handle numerical and categorical data and are easy to interpret and visualize. They work by creating a tree-like model of decisions based on the features of the data, with each internal node representing a decision based on one of the features and each leaf node representing a predicted class.
- Random Forests (Ho, 1995) are an ensemble method that combines multiple decision trees to make a more accurate and stable prediction. They work by training multiple decision trees on different subsets of the data and then aggregating their predictions through a process called majority voting.
- Support Vector Machines (Joachims, 1998) are a powerful and popular method for classification that is particularly effective for high-dimensional data. They work by finding the hyperplane in the feature space that maximally separates the different classes, and then using that hyperplane to classify new data points based on which side of the hyperplane they fall on.

These three algorithms were chosen because they are well-established, relatively easy to implement, fast, easily interpretable by humans, and have a track record of success in various NLP tasks (Locke, 2021), making them a good choice for evaluating the performance of different approaches to classifying stigmatizing and power/privilege language in clinical notes. We approached this NLP task as a binary classification where each note was assigned one label in each category (i.e., marginalizing language present vs. absent, and power/privilege language present vs. absent). All three machine learning algorithms were applied to generate classifications for each category (i.e., marginalizing language and

power/privilege language). This resulted in six machine learning models: three for marginalizing language and three for power/privilege language.

To evaluate the classification performance of the NLP methods, we used three standard metrics: precision, recall, and F-score. For the context of our study, "positive" denotes the presence of stigmatizing language. Precision represents the fraction of notes correctly identified as positive (based on manual review) among all notes that the model classified as positive. Recall measures the fraction of notes correctly identified as positive out of all notes that truly belong to the positive category (based on manual review). The F-score is the harmonic mean of precision and recall. These metrics range from 0, indicating poor performance, to 1, signaling perfect performance. NLP was implemented using the KNIME software (Berthold, 2009).

Finally, we applied a commonly used feature importance evaluation method (Information Gain [InfoGain]) to identify words that are important in classifying marginalizing and power/privilege language categories. InfoGain measures the reduction in entropy (degree of randomness) of a target variable after considering the values of a particular feature (Bridle, 1990). The feature with the highest InfoGain is considered to be the most informative and relevant to the target variable.

Results

The best-performing NLP methods achieved good classification performance on the test set for each category (Table 2). Decision Trees performed best for marginalizing language with an F-score of 0.73, while for power/privilege language, Support Vector Machines achieved an F-score of 0.91.

Table 2. Algorithm performance for detecting marginalizing and power/privilege language.

	Marginalizing Language			Power/privilege Language		
	Precision	Recall	F-Score	Precision	Recall	F-Score
Decision Trees	0.72	0.73	0.73	0.86	0.86	0.86
Random Forest	0.72	0.70	0.71	0.87	0.77	0.81
Support Vector Machines	0.63	0.65	0.64	0.88	0.95	0.91

In Table 3 we present the top 20 words that were most impactful in identifying marginalizing and power/privilege language categories. Words highly impactful in identifying marginalizing language often have negative connotations and appear in negative contexts. For example, there were several expressions hinting at the potential use or assessments for substance abuse, e.g., “illicit [substance]” and “marijuana.” Other words were related to potential language and immigration backgrounds (e.g., “bilingu[al],” “fluent [English],” “Dominican”), social determinants of health (e.g., “school,” “public [housing],” “apt [apartment],” “psychosoci[al]”), social work involvement (“LMSW [Licensed Master of Social Work]”), mental health assessments (“suicide [assessment]”) and unplanned pregnancy (“unplan[ned pregnancy]”).

On the other hand, words highly impactful in identifying power/privilege language tended to have positive connotations and appear in positive contexts. For example, several words referring to the hospitalized person’s appearance or emotional state were found, e.g., “[appropriately] dress[ed]/groom[ed],” “happy,” “friendly,” “cordial,” and “engaged.” Other language indicated social support (“[support system] compos[ed of]”) and employment status (employ[ed]).

Table 3: Top 20 words highly correlated with marginalizing and power/privilege language categories.

Marginalizing Language	Power/privilege language
illicit [substance]	nurture[ing marriage]
fluent	[support system] compos[ed of]
bilingu[al]	cordial
acknowledg[e]	[appropriately] groom[ed]
medicar [e]	[appropriately] dress[ed]
school	happy
file	pamphlet
public [housing]	psycho-educ [ation]
lmsw [licensed master of social work]	friend[ly]
suicide [assessment]	toward
await	engag[ed]
sourc [e]	strong
psychosoci[al]	family
live	marriag [e]
unknown	stressor
marijuana	incom [e]
apt	student
ssn	contribut [e]
dominican	employ[ed]
unplan[ned pregnancy]	list

Discussion

In this study, we identified well-performing NLP methods to automatically identify marginalizing language and power/privilege language in clinical notes. For identifying marginalizing language, the Decision Trees classifier performed best. The Support Vector Machines classifier was best at identifying power/privilege language with high F-score. We find these results promising and believe these methods serve as strong baselines for future studies of other NLP methods applied to this task. To our knowledge, this is the first study to use NLP performance metrics to evaluate the performance of NLP methods in identifying stigmatizing language. This is an important contribution to the field because previous studies have mostly relied on rule-based NLP approaches, such as keyword searches, to identify stigmatizing language in clinical notes. These previous approaches have the potential to produce false positives because they are based on a predetermined list of terms or expressions and may not accurately capture the full range of stigmatizing language used in clinical notes.

On the other hand, our machine learning approach is based on training a classifier using a labeled dataset of clinical notes that expert reviewers manually annotated. This allows the NLP methods to learn the contextual characteristics of marginalizing and power/privilege language and make predictions based on those characteristics. This approach is more accurate than rule-based approaches because it can capture the full range of stigmatizing language used in clinical notes and is less prone to producing false positives.

There are many clinical implications of this study. We found that machine learning algorithms can automatically detect stigmatizing language, and these algorithms have the potential to help to monitor clinical notes in real-time and identify instances of stigmatizing language. This can be especially relevant in the wake of the 21st Century Cures Act (United States Department of Health and Human Services, 2020), which is federal legislation passed in the United States to accelerate the development and approval of new medical treatments and technologies. One of the provisions of this Act is the requirement for EHR systems to provide individuals with timely access to their health information. There is some evidence to show that people may be offended by language used by clinicians in their clinical notes (Fernández et al., 2021), and this technology could provide resources for clinicians to address this issue. For example, machine learning algorithms could automatically detect stigmatizing language use and alert clinicians in real-time to remove it. This technology has the potential to improve the quality of care and the experience of hospitalized individuals by ensuring that they have access to information that is respectful and non-stigmatizing.

Further, our approach helped us to identify examples of language highly impactful in identifying the stigmatizing language in clinical notes. For example, clinical notes that contained marginalizing language were likely to have assessments for substance abuse. Since Black and Latinx birthing people are screened for substance abuse at higher rates than White people (Kravitz et al., 2021), this finding is concerning and requires further investigation. In addition, potential language barriers or immigration background were found to be associated with marginalizing language. Birthing people from immigrant communities may experience more barriers to care and receive a lower quality of care (Drewniak et al., 2017; Omenka et al., 2020); hence this finding also requires further examination in labor and birth settings. Finally, certain indicators of social determinants of health (e.g., income or public housing) were found to be correlated with marginalizing language in clinical notes. These findings highlight the need to address and eliminate language that may contribute to health disparities and perpetuate stigma. It is crucial for healthcare providers to be mindful of the language they use in medical notes and to avoid language that may reinforce negative stereotypes and contribute to health inequities.

Based on this work, there are implications for future research. Studies should be conducted to further develop and evaluate NLP methods for identifying stigmatizing language in clinical notes. For example, recent NLP methods, especially those based on deep neural networks, have shown promising results. Transformer-based language models utilizing self-attention (Vaswani, 2017) have nowadays become state-of-the-art across many NLP tasks. These are typically models that have first been pre-trained in a self-supervised manner on large text corpora before being fine-tuned on the task at hand (Devlin, 2019). We plan to explore the performance of such methods/models in detecting stigmatizing language in future research.

Another important area for future research is to explore whether clinicians disproportionately use language representing marginalized identities when referring to birthing people, including race and ethnicity. This is an important question because previous research has shown that some minoritized people have a higher risk of poor pregnancy and birth outcomes and may be more vulnerable to the negative effects of marginalizing language (Beach et al., 2021; Park et al., 2021). Further research is needed to understand whether marginalizing language is used disproportionately for these individuals and to identify the underlying causes of these disparities.

Finally, it will be important to examine the associations between stigmatizing language, quality of care, and health outcomes. For example, one recent study found that marginalizing language is associated with lower quality of pain management (Beach et al., 2021). Further research is needed to confirm and extend

these findings and to explore the mechanisms by which marginalizing language may affect outcomes for birthing people and their newborns.

There are also limitations to consider in interpreting the results of our study. First, our data come from two hospitals in urban settings in the northeastern United States and may not be generalizable to other hospitals or departments in different geographic settings. Second, we reviewed a relatively small sample of clinical notes, and it is possible that additional categories of stigmatizing language may be identified if a larger sample is reviewed. Furthermore, our employment of the Bag-of-Words (TF-IDF) model for text representation, while apt for our current dataset and scope, brings with it specific inherent challenges. The model, with its focus on word frequency, can sometimes obscure the more nuanced semantic meanings crucial for grasping the context or tone of documents. The sequence of words or semantic context, a pivotal aspect that can modify sentence meanings significantly, remains uncaptured by TF-IDF. Its limitations extend to discerning negations linked to words or phrases and to its sensitivity towards corpus size, potentially inflating the importance of infrequent words in smaller corpora. It also operates under the assumption of word independence, not accounting for potential word interactions or dependencies. Such limitations might impinge on the accuracy and robustness of our findings, especially when aiming to predict the presence of marginalizing and power/privilege language in clinical notes. To enhance future research undertakings, embracing advanced text representation techniques, such as word embeddings (e.g., Word2Vec, GloVe) and transformer-based models (e.g., BERT, GPT), could provide a richer insight into words' semantic and contextual depths. In addition, there remains a broader scope for refining our NLP methods, which could further boost the accuracy of our classifiers. Finally, this study's reliance on a traditional train-test split without employing cross-validation or bootstrapping might limit the comprehensiveness of our model's performance assessment.

In summary, our study highlights the importance of addressing stigmatizing language in clinical notes and provides a tool for identifying and addressing it. Further research is needed to confirm and extend these findings, particularly in other hospitals and departments and using larger datasets. This work has clinical implications for improving care for hospitalized people and has the potential to inform future research on the impact of stigmatizing language on quality of care and care outcomes.

Acknowledgements: This project was supported by funding from the Columbia University Data Science Institute Seeds Funds Program and a grant (GBMF9048) from the Gordon and Betty Moore Foundation.

Author contributions: Author contributions are as follows: Conceptualization (VB, MT), Analysis (DS, AD, BRI, MT), Original draft (VB), Revised draft (VB, DS, HM, AD, BRI, KC, MT), Funding (VB, MT, KC).

Declarations

Conflicts of interest: The authors have no conflicts of interest to disclose.

Human subjects: Human subjects approval for this study was received from the Institutional Review Board at Columbia Irving Medical Center, AAAT9870.

Data sharing: No new data were generated for this analysis, therefore, there are no data to share.

References

- Alpaydin, E. (2020). *Introduction to Machine Learning, fourth edition*. MIT Press. <https://books.google.com/books?id=tZnSDwAAQBAJ>
- Alpert, J. M., Morris, B. B., Thomson, M. D., Matin, K., Geyer, C. E., & Brown, R. F. (2019). OpenNotes in oncology: oncologists' perceptions and a baseline of the content and style of their clinician notes. *Transl Behav Med*, 9(2), 347-356. <https://doi.org/10.1093/tbm/iby029>
- Barcelona, V., Horton, R. L., Rivlin, K., Harkins, S., Green, C., Robinson, K., . . . Topaz, M. (2023). The Power of Language in Hospital Care for Pregnant and Birthing People: A Vision for Change. *Obstetrics & Gynecology*, 10.1097/AOG.0000000000005333. <https://doi.org/10.1097/aog.0000000000005333>
- Barcelona, V., Scharp, D., Ilday, B. R., Moen, H., Goffman, D., Cato, K., & Topaz, M. (2023). A qualitative analysis of stigmatizing language in birth admission clinical notes. *Nurs Inq*, e12557. <https://doi.org/10.1111/nin.12557>
- Beach, M. C., Saha, S., Park, J., Taylor, J., Drew, P., Plank, E., . . . Chee, B. (2021). Testimonial Injustice: Linguistic Bias in the Medical Records of Black Patients and Women. *Journal of general internal medicine*, 36(6), 1708-1714. <https://doi.org/10.1007/s11606-021-06682-z> [doi]
- Berthold, M. R. C., N.; Dill, F.; Gabriel, T.R.; Kotter, T.; Meinl, T.; Ohl, P.; Thiel, K.; Wiswedel, B. (2009). KNIME – The Konstanz Information Miner. *AcM SIGKDD explorations Newsletter*, 11(1), 26-31.
- Braveman, P., Dominguez, T. P., Burke, W., Dolan, S. M., Stevenson, D. K., Jackson, F. M., . . . Waddell, L. (2021). Explaining the Black-White Disparity in Preterm Birth: A Consensus Statement From a Multi-Disciplinary Scientific Work Group Convened by the March of Dimes [Review]. 3. <https://doi.org/10.3389/frph.2021.684207>
- Bridle, J. S. (1990). Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition. In F. F. Soulié, Hérault, J. (eds) (Ed.), *Neurocomputing* (Vol. 68). Berlin, Heidelberg.: Springer.
- Coyne, I. T. (1997). Sampling in qualitative research. Purposeful and theoretical sampling; merging or clear boundaries? *J Adv Nurs*, 26(3), 623-630. <https://doi.org/10.1046/j.1365-2648.1997.t01-25-00999.x>
- Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,
- Drewniak, D., Krones, T., & Wild, V. (2017). Do attitudes and behavior of health care professionals exacerbate health care disparities among immigrant and ethnic minority groups? An integrative literature review. *Int J Nurs Stud*, 70, 89-98. <https://doi.org/10.1016/j.ijnurstu.2017.02.015>
- Everett, B. G., Limburg, A., McKetta, S., & Hatzenbuehler, M. L. (2022). State-Level Regulations Regarding the Protection of Sexual Minorities and Birth Outcomes: Results From a Population-Based Cohort Study. *Psychosom Med*, 84(6), 658-668. <https://doi.org/10.1097/psy.0000000000001092>
- Fernández, L., Fossa, A., Dong, Z., Delbanco, T., Elmore, J., Fitzgerald, P., . . . DesRoches, C. (2021). Words Matter: What Do Patients Find Judgmental or Offensive in Outpatient Notes? *J Gen Intern Med*, 36(9), 2571-2578. <https://doi.org/10.1007/s11606-020-06432-7>
- Goddu, A. P., O'Connor, K. J., Lanzkron, S., Saheed, M. O., Saha, S., Peek, M. E., . . . Beach, M. C. (2018). Do Words Matter? Stigmatizing Language and the Transmission of Bias in the Medical Record. *Journal of general internal medicine*, 33(5), 685-691. <https://doi.org/10.1007/s11606-017-4289-2> [doi]
- Goh, Y. C., Cai, X. Q., Theseira, W., Ko, G., & Khor, K. A. (2020). Evaluating human versus machine learning performance in classifying research abstracts. *Scientometrics*, 125(2), 1197-1212. <https://doi.org/10.1007/s11192-020-03614-2>

- Hall, W. J., Chapman, M. V., Lee, K. M., Merino, Y. M., Thomas, T. W., Payne, B. K., . . . Coyne-Beasley, T. (2015). Implicit Racial/Ethnic Bias Among Health Care Professionals and Its Influence on Health Care Outcomes: A Systematic Review. *American Journal of Public Health*, 105(12), e60-76. <https://doi.org/10.2105/AJPH.2015.302903> [doi]
- Himmelstein, G., Bates, D., & Zhou, L. (2022). Examination of Stigmatizing Language in the Electronic Health Record. *JAMA Netw Open*, 5(1), e2144967. <https://doi.org/10.1001/jamanetworkopen.2021.44967>
- Ho, T. K. (1995). Random decision forests. . The Institute of Electrical and Electronics Engineers (IEEE), In Proceedings of 3rd international conference on document analysis and recognition
- Hoover, K., Lockhart, S., Callister, C., Holtrop, J. S., & Calcaterra, S. L. (2022). Experiences of stigma in hospitals with addiction consultation services: A qualitative analysis of patients' and hospital-based providers' perspectives. *J Subst Abuse Treat*, 138, 108708. <https://doi.org/10.1016/j.jsat.2021.108708>
- Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qual Health Res*, 15(9), 1277-1288. <https://doi.org/10.1177/1049732305276687>
- Jindal, M., Thornton, R. L. J., McRae, A., Unaka, N., Johnson, T. J., & Mistry, K. B. (2022). Effects of a Curriculum Addressing Racism on Pediatric Residents' Racial Biases and Empathy. *J Grad Med Educ*, 14(4), 407-413. <https://doi.org/10.4300/jgme-d-21-01048.1>
- Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. European conference on machine learning Berlin, Heidelberg.
- Kim, H., Sefcik, J. S., & Bradway, C. (2017). Characteristics of Qualitative Descriptive Studies: A Systematic Review. *Res Nurs Health*, 40(1), 23-42. <https://doi.org/10.1002/nur.21768>
- Kravitz, E., Suh, M., Russell, M., Ojeda, A., Levison, J., & McKinney, J. (2021). Screening for Substance Use Disorders during Pregnancy: A Decision at the Intersection of Racial and Reproductive Justice. *Am J Perinatol*. <https://doi.org/10.1055/s-0041-1739433>
- Landau, A. Y., Blanchard, A., Cato, K., Atkins, N., Salazar, S., Patton, D. U., & Topaz, M. (2022). Considerations for development of child abuse and neglect phenotype with implications for reduction of racial bias: a qualitative study. *J Am Med Inform Assoc*, 29(3), 512-519. <https://doi.org/10.1093/jamia/ocab275>
- Locke, S. B., A.; Al-Adely, S.; Moore, J.; Wilson, A.; Kitchen, G.B. (2021). Natural language processing in medicine: A review. *Trends in anaesthesia and critical care*, 38, 4-9. <https://doi.org/doi.org/10.1016/j.tacc.2021.02.007>
- Malouf, R., Redshaw, M., Kurinczuk, J. J., & Gray, R. (2014). Systematic review of health care interventions to improve outcomes for women with disability and their family during pregnancy, birth and postnatal period. *BMC Pregnancy Childbirth*, 14, 58. <https://doi.org/10.1186/1471-2393-14-58>
- Manning, C. D. R., P.; Schütze, H. (2008). *Introduction to information retrieval* (Vol. 39). Cambridge University Press.
- Martin, J. A., & Osterman, M. J. K. (2018). Describing the Increase in Preterm Births in the United States, 2014-2016. *NCHS data brief*, (312)(312), 1-8.
- Martin, K., & Stanford, C. (2020). An analysis of documentation language and word choice among forensic mental health nurses. *Int J Ment Health Nurs*, 29(6), 1241-1252. <https://doi.org/10.1111/inm.12763>
- Minehart, R. D., Bryant, A. S., Jackson, J., & Daly, J. L. (2021). Racial/Ethnic Inequities in Pregnancy-Related Morbidity and Mortality. *Obstet Gynecol Clin North Am*, 48(1), 31-51. <https://doi.org/10.1016/j.ogc.2020.11.005>

Omenka, O. I., Watson, D. P., & Hendrie, H. C. (2020). Understanding the healthcare experiences and needs of African immigrants in the United States: a scoping review. *BMC public health*, 20(1), 27.

<https://doi.org/10.1186/s12889-019-8127-9>

Park, J., Saha, S., Chee, B., Taylor, J., & Beach, M. C. (2021). Physician Use of Stigmatizing Language in Patient Medical Records. *JAMA network open*, 4(7).

<https://doi.org/10.1001/jamanetworkopen.2021.17052>

Philipsborn, R. P., Sorscher, E. A., Sexson, W., & Evans, H. H. (2021). Born on U.S. Soil: Access to Healthcare for Neonates of Non-Citizens. *Matern Child Health J*, 25(1), 9-14.

<https://doi.org/10.1007/s10995-020-03020-3>

Quinlan, J. R. (2014). C4. 5: Programs for Machine Learning. . 58-60.

https://books.google.com/books/about/C4_5.html?id=b3ujBQAAQBAJ

Sandelowski, M. (2010). What's in a name? Qualitative description revisited. *Res Nurs Health*, 33(1), 77-84. <https://doi.org/10.1002/nur.20362>

Shattell, M. M. (2009). Stigmatizing language with unintended meanings: "persons with mental illness" or "mentally ill persons"? *Issues Ment Health Nurs*, 30(3), 199. <https://doi.org/10.1080/01612840802694668>

Sun, M., Oliwa, T., Peek, M. E., & Tung, E. L. (2022). Negative Patient Descriptors: Documenting Racial Bias In The Electronic Health Record. *Health Aff (Millwood)*, 41(2), 203-211.

<https://doi.org/10.1377/hlthaff.2021.01423>

Tiway, U. S. S., T. (2008). *Natural Language Processing and Information Retrieval*. Oxford University Press, Inc. <https://dl.acm.org/doi/abs/10.5555/1481140>

Togioka, B. M., Seligman, K. M., & Delgado, C. M. (2022). Limited English proficiency in the labor and delivery unit. *Curr Opin Anaesthesiol*, 35(3), 285-291. <https://doi.org/10.1097/aco.0000000000001131>

United States Department of Health and Human Services. (2020, 08/04/2020). *21st Century Cures Act: Interoperability, information blocking, and the ONC health IT certification program*. National Archives. Retrieved November 5 from <https://www.federalregister.gov/documents/2020/05/01/2020-07419/21st-century-cures-act-interoperability-information-blocking-and-the-onc-health-it-certification>

Vaswani, A. S., N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. (2017). *Attention is all you need*. *Advances in neural information processing systems*.

<https://arxiv.org/abs/1706.03762>