Automatically Identifying Stigmatizing Language in Clinical Notes: A Survey of Natural Language Processing Methods

Annika Marie Schoene, Liz Scharnetzki, Jessica DiBiase, Althea Onaifo, Zara Poon, Tania Strout, Isha Agarwal

Submitted to: Journal of Medical Internet Research
on: October 22, 2024

Table of Contents

Automatically Identifying Stigmatizing Language in Clinical Notes: A Survey of Natural Language Processing Methods

Annika Marie Schoene[1*] PhD; Liz Scharnetzki[2*] PhD; Jessica DiBiase[2] MPH; Althea Onaifo[3]; Zara Poon[3]; Tania Strout[2, 4] PhD; Isha Agarwal[2, 4] PhD

[1]Institute for Experiential AI Northeastern University CHARLOTTE US
[2]Center for Interdisciplinary Population and Health Research MaineHealth Institute for Research Portland US
[3]Institute for Health Equity and Social Justice Research Northeastern University Boston US
[4]Department of Emergency Medicine Maine Medical Centre Portland US
[*]these authors contributed equally

Abstract

Background: Recording and communicating patient information is a vital part of providing care, however the (often unintentional) use of stigmatizing language in clinical care may contribute to health disparities among minoritized groups.

Objective: In this review, we aim to investigate the use of Natural Language Processing (NLP) to extract stigmatized language in clinical documentation, where we focus on social and cultural characteristics, types of stigma and NLP methods.

Methods: In this review, we follow PRISMA scoping review guidelines and extract metadata from each publication according to a predetermined set of categories that focus on (i) social and contextual characteristics associated with stigma (e.g., social identities, health status and condition), (ii) the type of stigmatizing language used, and (iii) Natural Language Processing (NLP) methods used to automatically identify stigma in EHRs.

Results: From our initial 1,882 papers we synthesize 13 papers that used methods grounded in Artificial Intelligence (AI) to automatically identify stigmatizing language in electronic health records (EHRs). We find that whilst stigmatizing language is analyzed in a variety of clinical settings, most studies primarily utilize on the use of only traditional machine learning algorithms. Furthermore, we find that most works used explicit word lists to extract stigmatizing language and focused on a limited set of social-identities and demographics.

Conclusions: We conclude that future research should focus on a boarder spectrum of social identities and demographics, developing methods that can also capture implicit measurements of stigma, and utilizing state-of-the-art language models and neural network architectures. More research is warranted to improve the detection of stigmatizing language by also incorporating intersectional definitions of stigma (identities, conditions) grounded in social psychological theories, investigating the contextual characteristics of the documentation (e.g.; provider, clinical environment), and ensuring NLP methods are reproducible and developed by interdisciplinary researchers.

(JMIR Preprints 22/10/2024:67807)
DOI: https://doi.org/10.2196/preprints.67807

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Automatically Identifying Stigmatizing Language in Clinical Notes: A Survey of Natural Language Processing Methods

Abstract

Background: Recording and communicating patient information is a vital part of providing care, however the (*often unintentional*) use of stigmatizing language in clinical care may contribute to health disparities among minoritized groups.
Objective: In this review, we aim to investigate the use of Natural Language Processing (NLP) to extract stigmatized language in clinical documentation, where we focus on social and cultural characteristics, types of stigma and NLP methods.
Methods: In this review, we follow PRISMA scoping review guidelines and extract metadata from each publication according to a predetermined set of categories that focus on (i) social and contextual characteristics associated with stigma (e.g., social identities, health status and condition), (ii) the type of stigmatizing language used, and (iii) Natural Language Processing (NLP) methods used to automatically identify stigma in EHRs.
Results: From our initial 1,882 papers we synthesize 13 papers that used methods grounded in Artificial Intelligence (AI) to automatically identify stigmatizing language in electronic health records (EHRs). We find that whilst stigmatizing language is analyzed in a variety of clinical settings, most studies primarily utilize on the use of only traditional machine learning algorithms. Furthermore, we find that most works used explicit word lists to extract stigmatizing language and focused on a limited set of social-identities and demographics.
Conclusion: We conclude that future research should focus on a boarder spectrum of social identities and demographics, developing methods that can also capture implicit measurements of stigma, and utilizing state-of-the-art language models and neural network architectures. More research is warranted to improve the detection of stigmatizing language by also incorporating intersectional definitions of stigma (identities, conditions) grounded in social psychological theories, investigating the contextual characteristics of the documentation (e.g.; provider, clinical environment), and ensuring NLP methods are reproducible and developed by interdisciplinary researchers.

Keywords: Healthcare; Natural Language Processing; Stigma; EHR; Machine Learning

Introduction

EHRs provide detailed sociodemographic and clinical information (e.g.; assigned triage acuity score, vital signs, laboratory and diagnostic imaging results), as well as free text notes that describe how a patient was evaluated and treated during specific clinical encounters. EHRs are essential for clinical staff in documenting, recommending, and communicating about the provision of care for their patients as they move through the healthcare system. While EHRs and the information they provide is essential, how this patient-level information is recorded and used can powerfully impact patient care. Action based on group-level patient information may give rise to the use of cognitive shortcuts, including stereotypes, in decision making. There is strong evidence, for example, of disparities in the delivery of care in a variety of clinical settings based on social identities (e.g., gender, race, ethnicity), socio-demographic characteristics (e.g., socioeconomic status, language) and, health conditions (e.g., diabetes, substance use disorder) [44]. Bias (i.e., preferential or skewed attitudes [73]) towards patients, either conscious or unconscious, has been identified as one of the primary drivers underlying disparities [59]; The presence of absence of stigmatizing language in the notes that providers write about patients, given that language, both spoken and written, may be a reflection of these negative biases.

Language, both spoken and written, is a fundamental way in which we share our mental constructs [74, 75]; thus, the continued use of stigmatizing language may be a mechanism for upholding social inequities. In the social sciences, bias is thought to be a type of skewed attitude or evaluation of an object, person, group, event, or idea [73, 76, 77]. Biases can be either positive or negative; implicit or explicit. While instrumental for rapid cognitive processing and decision making, when biases are applied to the characteristics of individuals or social groups, they can elicit problematic and consequential social sanctions such as social devaluation and stigmatization of specific social groups and personal attributes. In this work, we aim to operationalize stigma as a mark of disgrace associated with a particular social identity, circumstance, or characteristic that society has devalued (i.e., created a negative bias or attitude towards). In the context of language, words, phrases and sentiments that serve to devalue an individual on the basis of their identity, circumstance, or characteristic are, therefore, considered stigmatizing [60]. Existing studies and reviews in stigma research predominantly focus on qualitative approaches [13, 14, 32, 12,18, 26, 37, 19, 20, 21, 24, 27, 39] for specific groups of people [4,7,17, 28, 29], health conditions [5,6,7, 30, 31, 36] or clinical settings (e.g.: primary care [8]).

Natural Language Processing is an interdisciplinary subfield in Artificial Intelligence that refers to computational approaches (e.g.: machine learning, deep learning) to process, extract, and generate human language [78]. Previous work has applied NLP in medical settings to gain insight into the language used in EHRs at scale [33,34, 35]. Most recent work in AI and NLP investigating health disparities focuses on specific social identities that can be stigmatized (e.g.: race, ethnicity, gender, sexuality, etc.) [22, 23, 25, 38] or using other modalities (e.g.: video or audio recordings [16]). At the same time, there has been an increased focus in NLP to understand how datasets, learning models, and model outputs can be biased and lead to fairness concerns during practical applications of AI [41, 42, 43].

There are some overlaps in bias research in NLP (quantitative by nature) and qualitative stigma research in healthcare settings, where data types (e.g.: EHRs), methods (e.g.: the use and development of language models) and approaches (e.g.: a focus on certain sociodemographic groups that are disproportionally affected) are frequently used. However, the two fields can be distinct [40] in that (i) the words used to describe patient interactions and characteristics can only reveal their implied biased nature in certain contexts and people (e.g.: other clinical personnel that have access and experience in reading clinical records), (ii) unlike the authors in most NLP tasks, physicians are required to note down all patient interactions and describe in detail potentially socially devalued circumstances (e.g.: health or living conditions), and (iii) bias is differentially operationalized.

In this review, we follow PRISMA scoping review guidelines [80] and to the best of our knowledge, there is no existing review that investigates the use of NLP to extract stigmatized language in clinical documentation. To fill this gap in knowledge, we sought to answer the following research questions:

1. What are the social and contextual characteristics associated with the use of stigmatizing language that are being studied; specifically, what are the:
    a. Patient factors (e.g., social identities, health status and condition),
    b. Provider factors (e.g., profession, specialization), and
    c. Environmental factors (e.g., clinical setting) that are associated with use of stigmatizing language?
2. What type of stigmatizing language has been studied? (e.g.; stigmatizing language based on a patients' social identities versus health-based stigma)
    a. What definitions of stigma have been operationalized in the existing literature?
3. How is NLP used to detect stigma in clinical records? Specifically, what:
    a. datasets and preprocessing techniques,
    b. annotations and task formulations, and
    c. methods have been employed to date?

Methodology

Metadata Extraction

For each selected paper, we extracted the following metadata to identify trends and patterns in existing studies:

I. Social and Contextual Characteristics Associated with Stigma:

i Patient Factors: We captured basic demographic information (e.g.: age, sex and race) for people who either participated in the research study or who acted as a health care provider. In addition to this, we also included groups that are known to be disproportionately impacted by health disparities (e.g.: based on gender identity, sexuality, race and ethnicity) [9].

ii Provider Factors: Here, we identified the type of clinical provider involved (e.g., nurses, physicians, or surgeons).

iii Environmental Factors: Here, we identified the type of medical setting or facility the research study was conducted in (e.g., inpatient or outpatient).

II. Stigmatizing Language:

Here, we identified varying classifications of stigmatizing language being studied. Specifically, we created a distinction between these forms of stigmatization, such that devaluation of an individual on the basis of their social identity was considered *identity-based stigma* and devaluation of an individual on the basis of their health status or health condition was considered *health-based stigma*.

III. Natural Language Processing: In this category we extracted information about the type of NLP methods and tools used to identify stigma. This included, but was not limited to traditional machine learning approaches, such as topic modelling and Logistic Regression, but also neural network (e.g., Transformers) and rule-based approaches.

Selection Process

Search strategy and query: We conducted a search of scientific databases (Scopus, Web of Science, PubMed and the Anthology of the Association for Computational Linguistics) in May 2024, which captured published studies in medicine, computer science, NLP and the intersection of both fields. We included all papers that were peer-reviewed and published as a full text, between 2013 and 2024 using the following query:

(Stigma OR bias OR Identity OR Stereotype OR Disparity OR Belonging OR Exclusion) AND (Clinical Notes OR Electronic Health Records) AND (Language OR natural language processing OR text mining)

Filtering strategy. We identified 1*,882 papers* and after removing duplicates, retained *1,424 papers* in our two-step pre-screening process (see Figure 1). First, we identified a set of 12 seed papers, where 6 papers are closely related to our topic of interest and 6 are not relevant based on our inclusion and exclusion criteria (see full list of criteria below). Next, we utilized RobotAnalyst [1] to reduce human workload; RobotAnalyst is a popular open-source software tool that takes advantage of both machine learning and text mining techniques to categorize, cluster and rank collections of literature for their relevance (Free access to RobotAnalyst can be requested to reproduce this work). The tool is based on an iterative classification process and uses paper abstracts to decide if a paper is relevant. Furthermore, it enables researchers to select papers that should be included and excluded before classifying uncategorized papers. Here, we used our seed papers for inclusion and exclusion and RobotAnalyst returned *98 papers* after training and we retrained *67 papers* after screening for titles and abstracts. Finally, we manually screened all papers predicted to be included and retained *13 papers* for review, where we used the following inclusion and exclusion criteria:

Inclusion Criteria:

- Articles that utilized either text mining or NLP approaches to extract stigmatized language,
- Published studies that have been conducted in healthcare settings only,
- Articles that are published in English and focus on English-speaking countries, including USA, United Kingdom, Australia, and Ireland.

Exclusion Criteria:

- All abstracts, mission statements, commentary, review articles, and retractions;
- Articles that focused on bias in AI models or do not use language data to predict health disparities;
- Published studies that used qualitative methods to identify stigma in language;
- Articles that focused only on research involving human participants (e.g., observational, clinical trials).

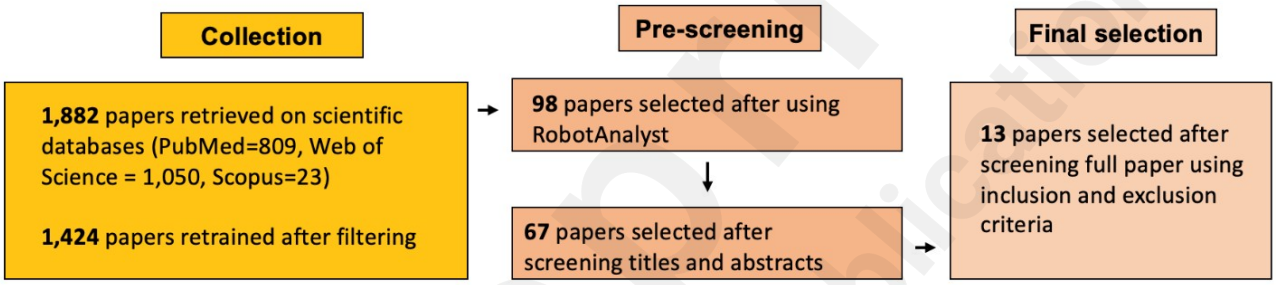

Figure 1: Overview of article screening and selection process.

Findings

Figure 2 depicts the number of papers published in our collection and whilst our search spanned over 10 years. We were only able to retrieve relevant literature since 2019 with no works published in 2020 and no relevant papers published before 2019. Furthermore, we anticipate that more literature on the subject will become available throughout 2024 and therefore is not reflected in our chart.

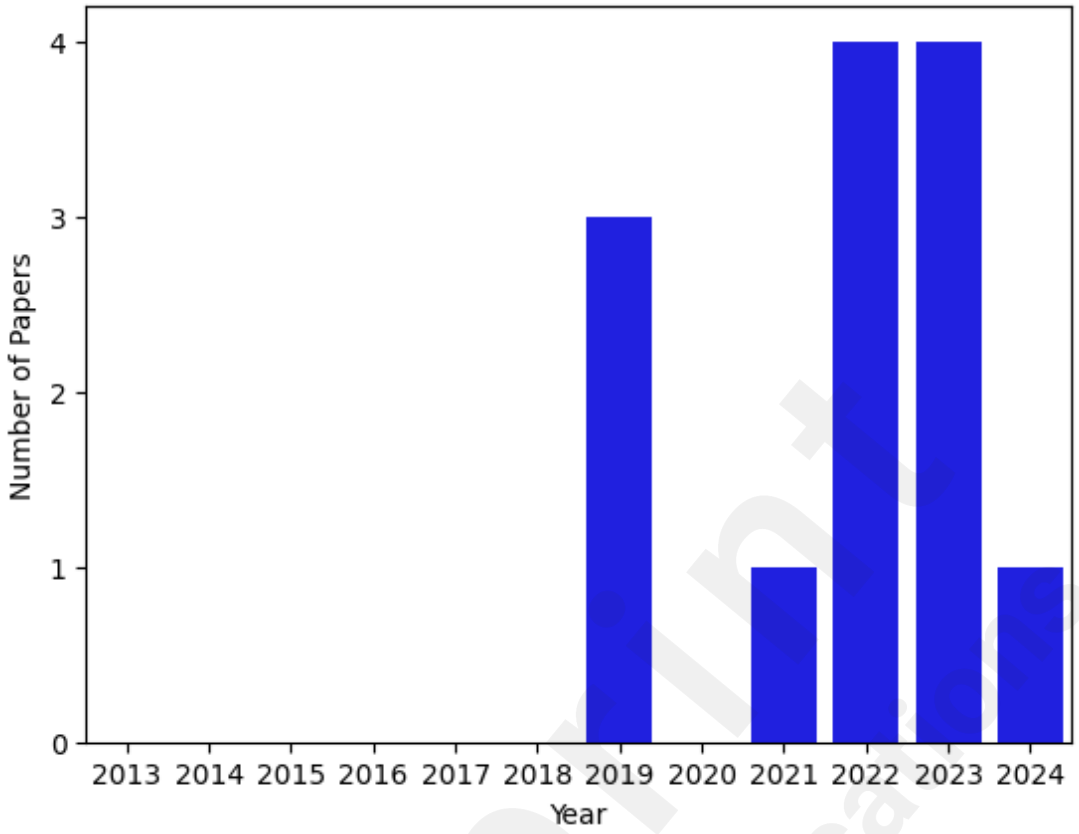

Figure 2: Number of papers published at the intersection of stigma and NLP until 2024.

Social and Contextual Characteristics Associated with Stigma

Patient Factors

In our collection of papers, most works account for multiple social identities [47, 28, 39, 19, 38, 46, 48, 21, 40], where only few studies focus on one *aspect of identity* such as racial descriptors [45, 49] or sexuality [25]. Frequently considered *aspects of identity* that can be *societally marginalized*, include race [47, 28, 39, 46, 21, 40] and ethnicity [39, 46, 21], where most papers focus on comparative approaches between Black [21, 39, 45, 46, 49], white [45, 46, 49, 39, 21], Hispanic [21, 39,46], and Asian [39, 46] patients. Gender identity [47, 48, 28, 38, 39], sex [19, 21, 40], and sexual orientation [25, 48] are other frequently considered factors. The majority of papers focus on a binary understanding of gender (e.g. male/female) [47, 28, 39, 21, 48] and only one paper looks at the impact of transgender people [48]. Less frequently investigated aspects of identity include age [47, 39, 19, 21], level of insurance coverage [47, 19], household income [47], marital status [19], and preferred language [47, 39].

Provider Factors

Work by several research groups [48, 49] and [19, 28, 47] focus on notes written by nurses and physicians, respectively. Other works in this sample analyze notes written by multiple different providers [38, 46, 21, 40, 27] from a variety of disciplines, including but not limited to social workers [38, 46, 21], nutritionists/dieticians [38, 27], and physical/occupational therapists [46]. Only one paper in our collection includes demographic information about the clinical providers [39] and five papers do not describe the type of clinical provider that authored the clinical note [25, 39, 45].

Environmental Factors

The most common clinical settings include Intensive Care Units (ICUs) [38, 45, 49, 40], home healthcare [46, 48], and emergency departments (EDs / ambulatory settings) [19, 28]. Two papers did not restrict themselves to one clinical setting, but focused on notes from multiple sources [21, 40], whereas other analyzed notes from a Veteran's Affairs clinic (VA) [25], pediatric hospital [47], and labor and birth admissions [27].

Health condition

The majority of papers do not explicitly outline the type of health condition experienced by the patients whose notes were analyzed [19, 28, 45, 46, 48, 49]. At the same time, some work focuses on multiple conditions at a time, such as septicemia and atherosclerosis [38] or diabetes, substance use, and chronic pain [39]. Few works look at specific health conditions, including substance use [21], epilepsy [47], HIV [25], and pregnancy [27].

Stigmatizing Language

The majority of papers included in this review [47, 25, 48, 28, 38, 49, 46, 45] focus on social-identity stigma [1, 3], however the types of analyzed social identities are limited to racial [40, 47], sexuality and gender [25, 48, 38, 40, 49, 46, 45]. Similarly, health / condition-based stigma is only evaluated for some conditions, such as substance use [21, 39] as well as chronic pain and diabetes [39]. Other types of stigmatizing language analyzed include negative patient descriptors [19] or marginalizing language [27].

Natural Language Processing

In the following section, we describe current datasets, tasks, common preprocessing steps, and algorithms used to identify stigma in clinical records.

Datasets and Preprocessing

The majority of articles in our survey either introduce their own dataset [40, 47, 25, 48, 19, 49, 46, 21, 27, 25], make use of freely available data [38, 49, 45, 40], such as subsets of MIMIC-III [50], a database of clinical records that includes both socio-demographic information and different types of notes among other variables. Most articles report the exact number of notes, unique patients and clinicians that make up their dataset where the smallest and largest dataset contain 1,117 [27] and over 110 million [45] notes, respectively. Typically, each note undergoes some form of preprocessing, a common and necessary step in developing NLP applications that involve the process of standardizing and normalizing textual data in preparation for further analysis and modelling [51]. For the papers in this survey, this usually involves tokenization using some form of pre-existing library [40, 28, 38] (e.g.: medspaCy [56] or spaCy [57]), [27, 19, 38, 40, 47] normalizing and lowercasing all tokens, [47, 48] removing stop words, punctuation and whitespaces, replacing numbers and specific keywords (e.g.: *non-compliant* to *non_compliant* [45]), as well as [45, 48, 39] generating n-grams using Gensim [55], and [28] adding Part-of-Speech (POS) tags [58]. Finally, all preprocessed notes are transformed into embedding representations [19, 49, 27, 45], such as TF-IDF [54] Word2Vec [53], and Bag-of-words (BOW) [52], that are used as input to a learning model.

Annotation and Task formulation

Annotation is a common step in NLP, and AI overall, that involves adding metadata and labels to unstructured data to enable any AI/ML algorithm to learn specific categories or characteristics [70]. Labels or metadata can be added at note, sentence or word-level in unstructured text by either humans or other algorithms (e.g.: using topic modelling algorithms to cluster similar documents and assign the topic category as a label to the overall document), depending on the task that needs to be

solved. For example, [40] first utilizes a dictionary containing stigmatizing words to identify relevant notes that were then manually labelled by two annotators. Similarly, [19, 27] also had expert annotators with training in health and data science label both sentences and parts of notes for (negative) patient descriptors. Most tasks reviewed in our articles were set up as classification tasks, where a note or sentence is automatically assigned into a predefined category. For example, work by [21, 39, 46] classify clinical notes into binary categories, including classifying notes with stigmatizing language based on a predetermined set of words. [27] uses two separate binary classifiers to identify stigma, where each model has to predict presence or absence of stigma, power or privilege. Recent studies setting up their task for multi-class classification, have labelled stigma using *positive* or *negative* sentiment and *out of context* [19] or *credibility/obstinacy*, *compliance* and *descriptors* [40]. Other work has extracted specific words [28], expanding on previous work by using NLP to detect new terms related to sexual orientation [25], and evaluating biases that can be carried over by NLP systems [49].

Methods

After completing all aforementioned steps of the NLP pipeline, each paper proceeded to either utilize a ML or rule-based algorithm, a neural network approach or NLP tools to classify notes or extract stigmatizing language depending on the task definition. The most popular approaches were based on traditional machine learning algorithms, including the use of Support Vector Machines (SVM) [47, 27], logistic regression [38, 39, 49], linear classifier with SGD [19], xgBoost [49], decision tree [27], and random forest [27]. Furthermore, readymade NLP tools were popular in multiple papers, such as FlashText [28], NimbleMiner [46], MTERMS NLP [21], KNIME [27], AutoMap [48], and SNOWMED-CT [25]. Rule-based approaches [21, 48] were also used for finding stigmatizing words, terms, phrases. Finally, two studies [40, 45] took advantage of more recent NLP methods by utilizing a Language Model (BERT [71]) and other neural network architectures that were not described in detail.

Discussion

It was the goal of the current review to examine (i) social and contextual characteristics associated with stigma, (ii) the type of stigmatizing language used, and (iii) Natural Language Processing (NLP) methods used to automatically identify stigma in EHRs. We were able to identify and synthesize 13 papers that showed promising results using NLP; however, several key challenges remain in automatically extracting stigmatizing language from EHRs. Here, we introduce open research questions and considerations for future work:

- *Social and Contextual Characteristics Associated with Stigma:* The basic demographic information assessed and analyzed in this collection is somewhat limited in that (i) there is often a focus on a single variable and (ii) works that look at multiple variables do not look at the possible compounding effect of belonging to more than one stigmatized demographic group (a concept known as intersectionality first conceived and popularized by Kimberlé Crenshaw [72]). Similarly, most work is solely focused on the differences between the treatment of white and Black patients, where very few look at other minoritized identities such as those who identify as Latino(a), Latine, Latinx or Hispanic and no work considers Native American populations. Whilst, this may in part be due to issues of such groups having access to the traditional healthcare system (leading to limited availability of information), it also means that most findings of words/concepts related to stigma are limited and likely not applicable / transferable to populations who are most stigmatized in the healthcare system. This leads us to ask the question: for whom we are developing new technologies and how broadly applicable they are [4]?

  There is also very little work focusing on the provider characteristics, which can lead

to a one-dimensional view of the overarching problem, where we only try to understand who is impacted and not who is often *unintentionally* at the root cause of the issue. Using this information could help to build more broadly applicable policies and interventions that improve patient care.

At the same time, environmental factors such as the clinical setting or a stigmatized health condition can play an important role in understanding how stigma shows up in real world settings. Especially for health-based stigma it is important to remember that people with different conditions may face different problems, where specific concepts are simply not transferable. Therefore, leading to the development of tools and technologies that can only be applied in a limited context. Overall, this leads us to believe that there is an increased need for more research that focuses on all sections of the patient and provider population as well as health conditions, so that any new tools or technologies developed capture a broader picture of the issue at large.

- *Stigmatizing Language:* In the case of both identity and health-based stigma, we found that the operationalization of bias and stigma were conflated. While it is true that in some instances, there are nuanced theoretical distinctions. Most operationaliztions of biased and stigmatizing language involves counting the number of words and phrases based on existing lexicons of terms and phrases. This strategy does not s not capture all possible experiences of stigma, as stigma is very often socially and contextually derived. Therefore, any lexicons or knowledge bases created for such work are inherently limited in themselves as they can only capture what they know. While it is important to capture explicit terminology, we cannot ignore the need for more broad measurements of stigma that may exist on an implicit spectrum. Similarly, the meaning of words and phrases is nuanced and context-dependent; in clinical documentation, the holistic meaning of the text may be more important than the use of a specific word or phrase. For example, standard triage documentation practices include the use of phrases such as “The patient refuses...” or “The patient complains”, but many lists include “refuse” or “complain” as a stigmatizing word. Alternatively, some words (e.g. “abuse”) may be stigmatizing in some contexts (e.g. when describing a patient with substance use disorder) but not in other contexts (e.g., when describing physical or sexual abuse that occurred towards a patient). Furthermore, even though category or group labels may be associated with stigma, the specific mechanisms by which those individuals are being devalued are not captured through this simple quantification of labels and word use. More research is needed focusing on the sentiments being evoked when potentially stigmatized social groups or conditions are being discussed to more holistically capture the experience of stigmatization.

- *Natural Language Processing:* Preprocessing of clinical data is a notoriously difficult task, where enough standardization and normalization is needed to reduce noise and build effective NLP models. However, sometimes the removal of certain punctuation (e.g., quotation marks as markers of implicit stigma), a common preprocessing step, can change the meaning, sentiment or intent of a note and therefore impact the NLP model’s ability to not just learn relevant features, but also make accurate new predictions over unseen data. At the same time, the methods used in this collection of papers are often limited in their description or can only be reproduced via the use of commercial (paid for) tools, which subsequently leads to a lack of reproducible and comparable research. Similarly, the majority of the datasets used are not available to the wider research community, which is often due to containing personal identifiable information (PII). Therefore, future research at the intersection of AI and healthcare should investigate and consider how data and NLP models can be shared responsibly to ensure that any scientific findings can be sufficiently peer-reviewed and

reproduced, ultimately leading to more transparent and applicable development of new technologies. This also applies to the dissemination of any high-quality human annotations, not only because the task in itself requires a lot of resources and is often costly as well as time-consuming, but also because high quality metadata and human knowledge are essential for developing AI/NLP models in critical settings (e.g.: healthcare or defense). Additionally, creating new tasks and tools in interdisciplinary settings can be challenging because rarely are teams diverse enough in their skillsets to either ground algorithms appropriately in clinical research, take advantage of state-of-the-art algorithms, or have sufficient and secure compute resources to run new NLP models. However, this presents new opportunities for future research where interdisciplinary collaborations between medical, social, and computer scientists could enable the application of newer methods such as Language Models, available through model zoos like Huggingface [79], or other neural network architectures (e.g., Transformer models [2]).

Conclusion

In this review, we have analyzed 13 scientific papers that used NLP to identify and extract stigmatizing language in EHRs. We found that most works used explicit word lists to extract stigmatizing language and focused on a limited set of social-identities and demographics. Findings suggest that current work is applying traditional machine learning techniques to detect and extract stigmatizing language, but rarely utilizes modern approaches. Finally, in our discussion we have highlighted both challenges and opportunities for future research, such as focusing on a boarder spectrum of social identities and demographics, developing methods that can also capture implicit measurements of stigma, and utilizing state-of-the-art language models and neural network architectures. Moreover, further research is warranted to improve the detection of stigmatizing language by also incorporating intersectional definitions of stigma (identities, conditions) grounded in social psychological theories, investigating the contextual characteristics of the documentation (e.g.; provider, clinical environment), and ensuring NLP methods are reproducible and developed by interdisciplinary researchers. The data supporting the findings of this study will be available upon request to the corresponding author.

Finally, this review has several limitations that qualify the generalizability of its findings. First, the lack of theoretical consensus on how stigma, bias, and general mechanisms of devaluation are being operationalized complicated the construction of inclusion criteria for this review. As a result, we were able to identify very few papers that accurately and consistently assessed these constructs. Secondly, we were unable to examine the intersectional nature of stigmas in this review given the generally limited focus on singular-identity based stigma. At the same time, this has limited the number of papers that have a stronger focus on NLP we have reviewed, where based on our definition of bias and stigma a large body of literature published at top venues (e.g.; annual conferences associated to the Association for Computational Linguistics or the Association for the advancement of artificial intelligence) did not qualify.

In spite of these limitations, this review highlights the need for future research to focus on not only a shared but broad operationalization of stigmatization, but also more modern machine learning approaches that are grounded in existing social science and psychology research.

Acknowledgements

A.M.S., A.O., and Z.P., searched the literature, and categorized each paper according to predefined categories. A.M.S generated visualizations. A.M.S. and L.S. wrote the initial draft and A.M.S., L.S.,

I.A. and T.S. revised the paper. All authors reviewed the paper.

Conflicts of Interest

None declared.

Abbreviations

AI: Artificial Intelligence
BOW: Bag-of-words
BERT: Bidirectional Encoder Representations from Transformers
EHR: Electronic Health Records
ML: Machine Learning
NLP: Natural Language Processing
POS: Part-of-Speech
SVM: Support Vector Machines: SVM
TF-IDF: Term frequency–inverse document frequency

References

[1] Goffman, E., 2014. Stigma and social identity. In Understanding deviance (pp. 256-265). Routledge.
[2] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M. and Davison, J., 2020, October. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations (pp. 38-45).

[3] Major, B. and O'brien, L.T., 2005. The social psychology of stigma. Annu. Rev. Psychol., 56, pp.393-421.
[4] Harari, L. and Lee, C., 2021. Intersectionality in quantitative health disparities research: A systematic review of challenges and limitations in empirical studies. Social science & medicine, 277, p.113876.
[5] Katz, I.T., Ryu, A.E., Onuegbu, A.G., Psaros, C., Weiser, S.D., Bangsberg, D.R. and Tsai, A.C., 2013. Impact of HIV-related stigma on treatment adherence: systematic review and meta-synthesis. Journal of the International AIDS Society, 16, p.18640.
[6] Unit, C.E., Racial discrimination and adverse pregnancy outcomes: a systematic review and meta-analysis.
[7] Place, V., Nabb, B., Viksten Assel, K., Bäärnhielm, S., Dalman, C. and Hollander, A.C., 2021. Interventions to increase migrants' care-seeking behaviour for stigmatised conditions: a scoping review. Social psychiatry and psychiatric epidemiology, 56, pp.913-930.
[8] Wang, J.X., Somani, S., Chen, J.H., Murray, S. and Sarkar, U., 2021. Health equity in artificial intelligence and primary care research: Protocol for a scoping review. JMIR research protocols, 10(9), p.e27799.
[9] National Institutes of Health, Minority Health and Health Disparities: Definitions and Parameters.
[10] Eng, K., Johnston, K., Cerda, I., Kadakia, K., Mosier-Mills, A. and Vanka, A., 2024. A Patient-Centered Documentation Skills Curriculum for Preclerkship Medical Students in an Open Notes Era. MedEdPORTAL, 20, p.11392.
[11] Korach, Z.T., Yang, J., Rossetti, S.C., Cato, K.D., Kang, M.J., Knaplund, C., Schnock, K.O., Garcia, J.P., Jia, H., Schwartz, J.M. and Zhou, L., 2020. Mining clinical phrases from nursing notes to discover risk factors of patient deterioration. International journal of medical informatics, 135, p.104053.
[12] Roy, M., Purington, N., Liu, M., Blayney, D.W., Kurian, A.W. and Schapira, L., 2021. Limited

English proficiency and disparities in health care engagement among patients with breast cancer. JCO Oncology Practice, 17(12), pp.e1837-e1845.
[13] Thom, R.P. and Farrell, H.M., 2017. When and how should clinicians share details from a health record with patients with mental illness?. AMA Journal of Ethics, 19(3), pp.253-259.
[14] Bell, S.K., Mejilla, R., Anselmo, M., Darer, J.D., Elmore, J.G., Leveille, S., Ngo, L., Ralston, J.D., Delbanco, T. and Walker, J., 2017. When doctors share visit notes with patients: a study of patient and doctor perceptions of documentation errors, safety opportunities and the patient–doctor relationship. BMJ quality & safety, 26(4), pp.262-270.
[15] Sterling, N.W., Patzer, R.E., Di, M. and Schrager, J.D., 2019. Prediction of emergency department patient disposition based on natural language processing of triage notes. International journal of medical informatics, 129, pp.184-188.
[16] Hsueh, L., Huang, J., Millman, A.K., Gopalan, A., Parikh, R.K., Teran, S. and Reed, M.E., 2023. Cross-Sectional Association of Patient Language and Patient-Provider Language Concordance with Video Telemedicine Use Among Patients with Limited English Proficiency. Journal of General Internal Medicine, 38(3), pp.633-640.
[17] Pelleboer-Gunnink, H.A., Van Oorsouw, W.M.W.J., Van Weeghel, J. and Embregts, P.J.C.M., 2017. Mainstream health professionals' stigmatising attitudes towards people with intellectual disabilities: a systematic review. Journal of Intellectual Disability Research, 61(5), pp.411-434.
[18] Pérez-Stable, E.J. and El-Toukhy, S., 2018. Communicating with diverse patients: how patient and clinician factors affect disparities. Patient education and counseling, 101(12), pp.2186-2194.
[19] Sun, M., Oliwa, T., Peek, M.E. and Tung, E.L., 2022. Negative Patient Descriptors: Documenting Racial Bias In The Electronic Health Record: Study examines racial bias in the patient descriptors used in the electronic health record. Health Affairs, 41(2), pp.203-211.
[20] P Goddu, A., O'Conor, K.J., Lanzkron, S., Saheed, M.O., Saha, S., Peek, M.E., Haywood, C. and Beach, M.C., 2018. Do words matter? Stigmatizing language and the transmission of bias in the medical record. Journal of general internal medicine, 33, pp.685-691.
[21] Weiner, S.G., Lo, Y.C., Carroll, A.D., Zhou, L., Ngo, A., Hathaway, D.B., Rodriguez, C.P. and Wakeman, S.E., 2023. The incidence and disparities in use of stigmatizing language in clinical notes for patients with substance use disorder. Journal of addiction medicine, pp.10-1097.
[22] Carreras Tartak, J.A., Brisbon, N., Wilkie, S., Sequist, T.D., Aisiku, I.P., Raja, A. and Macias-Konstantopoulos, W.L., 2021. Racial and ethnic disparities in emergency department restraint use: a multicenter retrospective analysis. Academic Emergency Medicine, 28(9), pp.957-965.
[23] Wieland, M.L., Wu, S.T., Kaggal, V.C. and Yawn, B.P., 2013. Tracking health disparities through natural-language processing. American journal of public health, 103(3), pp.448-449.
[24] Park, J., Saha, S., Chee, B., Taylor, J. and Beach, M.C., 2021. Physician use of stigmatizing language in patient medical records. JAMA Network Open, 4(7), pp.e2117052-e2117052.
[25] Lynch, K.E., Alba, P.R., Viernes, B. and DuVall, S.L., 2019, August. Using enriched samples for semi-automated vocabulary expansion to identify rare events in clinical text: sexual orientation as a use case. In MedInfo (pp. 1532-1533).
[26] Roy, M., Purington, N., Liu, M., Blayney, D.W., Kurian, A.W. and Schapira, L., 2021. Limited English proficiency and disparities in health care engagement among patients with breast cancer. JCO Oncology Practice, 17(12), pp.e1837-e1845.
[27] Barcelona, V., Scharp, D., Idnay, B.R., Moen, H., Goffman, D., Cato, K. and Topaz, M., 2023. A qualitative analysis of stigmatizing language in birth admission clinical notes. Nursing Inquiry, p.e12557.
[28] Beach, M.C., Saha, S., Park, J., Taylor, J., Drew, P., Plank, E., Cooper, L.A. and Chee, B., 2021. Testimonial injustice: linguistic bias in the medical records of black patients and women. Journal of general internal medicine, 36(6), pp.1708-1714.
[29] Cruz, T.M. and Smith, S.A., 2021. Health equity beyond data: health care worker perceptions of race, ethnicity, and language data collection in electronic health records. Medical Care, 59(5),

pp.379-385.
[30] Kaufmann, J., Marino, M., Lucas, J.A., Rodriguez, C.J., Bailey, S.R., April-Sanders, A.K., Boston, D. and Heintzman, J., 2022. Racial, ethnic, and language differences in screening measures for statin therapy following a major guideline change. Preventive medicine, 164, p.107338.
[31] Witting, C., Azizi, Z., Gomez, S.E., Zammit, A., Sarraju, A., Ngo, S., Hernandez-Boussard, T. and Rodriguez, F., 2023. Natural language processing to identify reasons for sex disparity in statin prescriptions. American Journal of Preventive Cardiology, 14, p.100496.
[32] Fernández, L., Fossa, A., Dong, Z., Delbanco, T., Elmore, J., Fitzgerald, P., Harcourt, K., Perez, J., Walker, J. and DesRoches, C., 2021. Words matter: what do patients find judgmental or offensive in outpatient notes?. Journal of general internal medicine, pp.1-8.
[33] Fanconi, C., van Buchem, M. and Hernandez-Boussard, T., 2023. Natural Language Processing Methods to Identify Oncology Patients at High Risk for Acute Care with Clinical Notes. AMIA Summits on Translational Science Proceedings, 2023, p.138.
[34] Cohen, R., Elhadad, M. and Elhadad, N., 2013. Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies. BMC bioinformatics, 14(1), pp.1-15.
[35] Song, J., Ojo, M., Bowles, K.H., McDonald, M.V., Cato, K., Rossetti, S.C., Adams, V., Chae, S., Hobensack, M., Kennedy, E. and Tark, A., 2022. Detecting language associated with home healthcare patient's risk for hospitalization and emergency department visit. Nursing research, 71(4), pp.285-294.
[36] Martin, K. and Stanford, C., 2020. An analysis of documentation language and word choice among forensic mental health nurses. International Journal of Mental Health Nursing, 29(6), pp.1241-1252.
[37] Gill, M., Cohen-Cline, H., Holtorf, M. and Vartanian, K., 2023. Mammogram perceptions, communication, and gaps in care among individuals with non-English language preference in Oregon and Washington states. Preventive Medicine Reports, 35, p.102352.
[38] Penn, J.A. and Newman-Griffis, D., 2022. Half the picture: Word frequencies reveal racial differences in clinical documentation, but not their causes. In AMIA Annual Symposium Proceedings (Vol. 2022, p. 386). American Medical Informatics Association.
[39] Himmelstein, G., Bates, D. and Zhou, L., 2022. Examination of stigmatizing language in the electronic health record. JAMA Network Open, 5(1), pp.e2144967-e2144967.
[40] Harrigian, K., Zirikly, A., Chee, B., Ahmad, A., Links, A., Saha, S., Beach, M.C. and Dredze, M., 2023, July. Characterization of stigmatizing language in medical records. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 312-329).
[41] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A., 2021. A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6), pp.1-35.
[42] Pessach, D. and Shmueli, E., 2022. A review on fairness in machine learning. ACM Computing Surveys (CSUR), 55(3), pp.1-44.
[43] Norori, N., Hu, Q., Aellen, F.M., Faraci, F.D. and Tzovara, A., 2021. Addressing bias in big data and AI for health care: A call for open science. Patterns, 2(10).
[44] Puissant MM, Agarwal I, Scharnetzki E, Cutler A, Gunnell H, Strout TD. Racial differences in triage assessment at rural vs urban Maine emergency departments. Intern Emerg Med. Published online April 10, 2024. doi:10.1007/s11739-024-03560-4
[45] Cobert, J., Mills, H., Lee, A., Gologorskaya, O., Espejo, E., Jeon, S.Y., Boscardin, W.J., Heintz, T.A., Kennedy, C.J., Ashana, D.C. and Chapman, A.C., 2024. Measuring Implicit Bias in ICU Notes Using Word-Embedding Neural Network Models. Chest.
[46] Topaz, M., Song, J., Davoudi, A., McDonald, M., Taylor, J., Sittig, S. and Bowles, K., 2023. Home Health Care Clinicians' Use of Judgment Language for Black and Hispanic Patients: Natural Language Processing Study. JMIR nursing, 6, p.e42552.

[47] Wissel, B.D., Greiner, H.M., Glauser, T.A., Mangano, F.T., Santel, D., Pestian, J.P., Szczesniak, R.D. and Dexheimer, J.W., 2019. Investigation of bias in an epilepsy machine learning algorithm trained on physician notes. Epilepsia, 60(9), pp.e93-e98.
[48] Bjarnadottir, R.I., Bockting, W., Yoon, S. and Dowding, D.W., 2019. Nurse documentation of sexual orientation and gender identity in home healthcare: a text mining study. CIN: Computers, Informatics, Nursing, 37(4), pp.213-221.
[49] Adam, H., Yang, M.Y., Cato, K., Baldini, I., Senteio, C., Celi, L.A., Zeng, J., Singh, M. and Ghassemi, M., 2022, July. Write it like you see it: Detectable differences in clinical notes by race lead to differential model recommendations. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (pp. 7-21).
[50] Johnson, A., Pollard, T., & Mark, R. (2016). MIMIC-III Clinical Database (version 1.4). PhysioNet. https://doi.org/10.13026/C2XW26.
[51] Loper, E. and Bird, S., 2002, July. NLTK: the Natural Language Toolkit. In Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1(pp. 63-70).
[52] Zhang, Y., Jin, R. and Zhou, Z.H., 2010. Understanding bag-of-words model: a statistical framework. International journal of machine learning and cybernetics, 1, pp.43-52.
[53] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J., 2013. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26.
[54] Sparck Jones, K., 1972. A statistical interpretation of term specificity and its application in retrieval. Journal of documentation, 28(1), pp.11-21.
[55] Řehůřek, R. and Sojka, P., 2011. Gensim—statistical semantics in python. Retrieved from genism.org.
[56] Eyre, H., Chapman, A.B., Peterson, K.S., Shi, J., Alba, P.R., Jones, M.M., Box, T.L., DuVall, S.L. and Patterson, O.V., 2021. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. In AMIA Annual Symposium Proceedings (Vol. 2021, p. 438). American Medical Informatics Association.
[57] Honnibal, M. & Montani, I., 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.
[58] Manning, C.D., 2011, February. Part-of-speech tagging from 97% to 100%: is it time for some linguistics?. In International conference on intelligent text processing and computational linguistics (pp. 171-189). Berlin, Heidelberg: Springer Berlin Heidelberg.
[59] Hatzenbuehler, M. L., Phelan, J. C., & Link, B. G. (2013). Stigma as a fundamental cause of population health inequalities. American journal of public health, 103(5), 813–821. https://doi.org/10.2105/AJPH.2012.301069
[60] Major, B., & O'Brien, L. T. (2005). The social psychology of stigma. Annual review of psychology, 56, 393–421. https://doi.org/10.1146/annurev.psych.56.091103.070137
[70] Alpaydin, E., 2020. Introduction to machine learning. MIT press.
[71] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2019, June. 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, Volume 1 (Long and Short Papers) (pp. 4171-4186).
[72] Crenshaw, K., 2013. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In Feminist legal theories (pp. 23-51). Routledge.
[73] VandenBos, G.R., 2007. APA dictionary of psychology. American Psychological Association.
[74] Boroditsky, L., 2011. How language shapes thought. Scientific American, 304(2), pp.62-65.sa
[75] Whorf, B.L., 2012. Language, thought, and reality: Selected writings of Benjamin Lee Whorf. MIT press.

[76] Chen, M. and Bargh, J.A., 1997. Nonconscious behavioral confirmation processes: The self-fulfilling consequences of automatic stereotype activation. Journal of Experimental Social Psychology, 33(5), pp.541-560.
[77] Dovidio, J.F., Kawakami, K. and Beach, K.R., 2001. Implicit and explicit attitudes: Examination of the relationship between measures of intergroup bias. Blackwell handbook of social psychology: Intergroup processes, 4, pp.175-197.
[78] Manning, C.D., 2022. Human language understanding & reasoning. Daedalus, 151(2), pp.127-138.
[79] Wolf, T., 2019. Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.
[80] Tricco, A.C., Lillie, E., Zarin, W., O'Brien, K.K., Colquhoun, H., Levac, D., Moher, D., Peters, M.D., Horsley, T., Weeks, L. and Hempel, S., 2018. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Annals of internal medicine*, *169*(7), pp.467-473.