JAMIA Open, 5(2), 2022, 1–11
https://doi.org/10.1093/jamiaopen/ooac039
Review

Review

A scoping review of ethics considerations in clinical natural language processing

Oliver J. Bear Don't Walk IV[1], Harry Reyes Nieva [1,2], Sandra Soo-Jin Lee[3], and Noémie Elhadad[1]

[1]Department of Biomedical Informatics, Columbia University, New York, New York, USA, [2]Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA, and [3]Department of Medical Humanities and Ethics, Columbia University, New York, New York, USA

Corresponding Author: Oliver J. Bear, MA, Don't Walk IV, Department of Biomedical Informatics, Columbia University, 622 West 168th Street, PH20, New York, NY 10032, USA; ob2285@cumc.columbia.edu

Received 24 January 2022; Revised 5 May 2022; Editorial Decision 8 May 2022; Accepted 12 May 2022

ABSTRACT

Objectives: To review through an ethics lens the state of research in clinical natural language processing (NLP) for the study of bias and fairness, and to identify gaps in research.
Methods: We queried PubMed and Google Scholar for articles published between 2015 and 2021 concerning clinical NLP, bias, and fairness. We analyzed articles using a framework that combines the machine learning (ML) development process (ie, design, data, algorithm, and critique) and bioethical concepts of beneficence, nonmaleficence, autonomy, justice, as well as explicability. Our approach further differentiated between biases of clinical text (eg, systemic or personal biases in clinical documentation towards patients) and biases in NLP applications.
Results: Out of 1162 articles screened, 22 met criteria for full text review. We categorized articles based on the design ($N = 2$), data ($N = 12$), algorithm ($N = 14$), and critique ($N = 17$) phases of the ML development process.
Discussion: Clinical NLP can be used to study bias in applications reliant on clinical text data as well as explore biases in the healthcare setting. We identify 3 areas of active research that require unique ethical considerations about the potential for clinical NLP to address and/or perpetuate bias: (1) selecting metrics that interrogate bias in models; (2) opportunities and risks of identifying sensitive patient attributes; and (3) best practices in reconciling individual autonomy, leveraging patient data, and inferring and manipulating sensitive information of subgroups. Finally, we address the limitations of current ethical frameworks to fully address concerns of justice. Clinical NLP is a rapidly advancing field, and assessing current approaches against ethical considerations can help the discipline use clinical NLP to explore both healthcare biases and equitable NLP applications.

Key words: natural language processing, bias, fairness, ethically informed

LAY SUMMARY

The objective of this work is to explore the ethical considerations of clinical natural language processing (NLP) in the context of bias. Bias here refers to systematic differences in terms of representation or application of NLP models between group identities like race, gender, and sexuality. We searched PubMed and Google Scholar for articles concerning NLP, ethics, and bias between 2015 and 2021. We analyzed articles against a framework that combines different stages of the machine learning (ML) development process (design, data, algorithm, critique) and important ethical principles in the medical domain. We included 22 out of 1162 prescreened articles in this review. Articles were categorized into: design ($N=2$), data ($N=12$), algorithm ($N=14$), and critique ($N=17$). Clinical NLP can be used to study bias in research that relies on clinical text as well as explore biases in the healthcare setting. We identify 3 areas of active research at the intersection of clinical NLP and ethics: (1) selecting performance metrics that interrogate bias in ML; (2) opportunities and risks of identifying sensitive patient information like gender, and sexuality; and (3) best practices in balancing individual autonomy, leveraging patient data, and inferring and manipulating sensitive information of subgroups.

INTRODUCTION

Recently, there has been a sharp increase in research at the intersection of machine learning (ML) and bias in clinical and biomedical research.[1–6] While ML approaches may help advance human health, they also hold the potential to further entrench existing, even well-documented, biases. Such biases have led to growing concerns regarding the exacerbation of healthcare disparities,[7] lack of funding for certain research topics,[8] and inadequate diversity in the demographic makeup of study populations.[9] Prior work has largely focused on elucidating biases via quantitative analysis of structured electronic health record (EHR) data.[6,10,11] Natural language processing (NLP) of clinical text presents yet another robust method for inquiry, and is differentiated from general ML as methods or algorithms that take in or produce unstructured, free-text data.[12] Most notably, NLP-based work has uncovered gender differences in disease associations,[3] disparities in smoking documentation,[13] and differences in financial consideration discussions[14] among racial and ethnic groups. These examples offer a glimpse into biases in both the practice and discussion of healthcare delivery. Beyond discovery of bias in clinical text, NLP may also offer solutions (eg, techniques for information extraction have improved identification and representation of subgroups in clinical data[15–18]), though, unless actively monitored and addressed, clinical NLP can contribute to biases through its development and use.

Clinical text is a rich and nuanced source of patient data, but the subjective and nonstandardized means by which information is recorded and discussed in clinical notes also raises unique ethical considerations. Moreover, NLP has been widely leveraged in a multitude of tasks (eg, information extraction,[15–22] understanding clinical workflow,[21,23] risk prediction and patient stratification,[24–26] patient trajectory prediction,[27,28] decision support,[29] and question answering[30]) without a structured approach to examining and understanding the sources and implications of its biases in the complex environment of clinical practice. Multiple agents work to maximize benefits, minimize harms, and respect autonomy while maintaining a fair distribution of resources within the biomedical ecosystem. The 4 core bioethical principles of beneficence, nonmaleficence, autonomy, and justice present a framework for ethical decision making that allows for the complexity of healthcare delivery. Nonetheless, to fully examine and understand bias and its relevance to the application of NLP in the clinical setting, an expanded framework is necessary.

Ethical concerns surrounding ML have been discussed by many researchers, in particular bias and fairness.[31–34] Related studies have demonstrated that ML models can inherit, exacerbate, or even create new biases leading to disparities.[31] Similar to bioethics, this work involves multiple agents interacting within various environments.[35] However, to our knowledge, focused study of bias and ethics of clinical NLP remains a relatively nascent domain. Given the growing body of literature concerning ethical ML and the sensitive nature of clinical text, it is important to understand and anticipate ethical concerns before NLP applications are put into practice.

A scoping review is well suited to provide an overview of bias in clinical NLP as they can rapidly map key concepts "underpinning a research area."[36] The objective of this work was to perform a scoping review of literature at the intersection between clinical NLP, bias, and fairness. It incorporates a robust framework that combines traditional bioethical principles with the stages of a proposed ML development process. This approach offers a unique lens through which clinical NLP and its broader ethical implications on healthcare decision making may be viewed and better understood. Overall, we find that clinical NLP can be used to uncover and ameliorate bias in healthcare, but is not without its own ethical concerns and even well-intentioned work can potentially expose patients to harm. While clinical NLP can support research into biases in the clinical setting, it also has the potential to inherit or exacerbate the biases we hope to study, or even lead to new biases altogether.[37]

METHODS

We conducted our scoping review based on recommendations from the PRISMA Extension for Scoping Reviews (PRISMA-ScR) guidelines.[38]

Eligibility criteria

We included 2 types of articles: (1) empirical studies on identifying or mitigating bias in clinical notes and (2) tasks focused on predictive analytics, classification, or information extraction using clinical text. We excluded articles if they did not focus on English text or did not involve a data-driven approach. Additionally, we further required articles to measure bias. Of note, our definition of bias differs from the term bias in ML literature that describes the differences between an estimator's expected value and the true value of the parameter being estimated.[39] In this study, we define bias in a sociological sense, that is still tied to machine learning. We defined bias as systematic differences in representations, predictions, or outcomes for individuals correlated with inherent or acquired characteristics related to systemic marginalization.

Search strategy

We searched PubMed and Google Scholar on June 8, 2021 using search terms related to the concepts of NLP, clinical data, and bias

Table 1. Search terms and queries for PubMed and Google Scholar

Source	Search term			Query
	NLP	Clinical data	Bias	
PubMed	"natural language processing", "machine learning", "artificial intelligence", "information storage and retrieval"	"unstructured", "electronic health records", "clinical"	"bias*", "fair", "fairness", "health disparities", "explicability", "interpretab*", "explainab*"	("natural language processing" OR "machine learning" OR "artificial intelligence" OR "information storage and retrieval") AND ("unstructured" OR "electronic health records" OR "clinical") AND ("bias*" OR "fair" OR "fairness" OR "health disparities" OR "explicability" OR "interpretab*" OR "explainab*")
Google Scholar	"natural language processing", "machine learning"	"clinical note", "clinical text", "electronic health records"	"bias*", "fairness", "health disparities"	("natural language processing" OR ("machine learning" AND ("clinical note" OR "clinical text"))) AND ("electronic health records") AND ("bias*" OR "fairness" OR "health disparities"

(Table 1) and included all work published since 2015 (inclusive). To account for concepts related to bias that fall under our definition we expanded our search to include terms such as fairness, health disparities, and explicability. We focused our review on more recent discussions of bias in clinical NLP as the ML and NLP-related work has been advancing at a rapid pace.

Study selection

Two reviewers (OJBDW and HRN) independently screened all titles and abstracts to determine eligibility for full text review. Any disagreements were adjudicated to reach consensus and an additional reviewer (SS-JL) was consulted, if needed.

Data extraction

Briefly, we analyzed articles along 2 axes, ML and ethics, to capture which aspects of both fields were investigated thus far in the literature. The ML axis (Figure 1) was modeled after existing work on ethical ML in healthcare[40] and Box's Loop[41] and serves to conceptualize the stages of developing ML. For this component, we analyzed the 4 phases of each paper: study design (which is influenced by the research objectives), choice of data source (and possible selection biases), algorithm employed (and the features used for algorithm input), and self-critique (with respect to aspects such as assumptions made and the relative importance of individual elements of the work). The ethics axis (Figure 2) is based on a synthesis of existing ethical frameworks conducted by AI4People, a multi-stakeholder forum concerned with laying the foundations for a "Good AI Society."[42] Each dimension (beneficence, nonmaleficence, justice, autonomy, and explicability) was tailored for its application to ML development and adoption.

The ML axis integrates the 4 stages of the ML development process introduced by Chen et al.[40]—design, data, algorithm, and critique. For the critique stage, we also incorporated a cyclic graph structure as a generalization of the principles of Box's Loop,[41] which proposes that building and computing probabilistic model is an iterative process (we extend this definition to all ML models). The ML development process begins in the design stage during which researchers determine what is being studied and which stakeholders are included. During the data stage, decisions are made concerning how to collect data, inclusion and exclusion criteria, what features to extract, and which groups are studied and how they are defined. Next, the algorithm stage refers to algorithm choice, training, optimization, and validation. In this phase, researchers may choose to optimize for a proxy if the true outcome is too difficult to measure.[10,43] Examples of proxy labels in NLP tasks include: predicting missing words and logical sentence flow[44] or billing codes[45] to learn general language representations, and quality of life as measured by standardized surveys.[46] The critique stage is unique in that it interacts with all other stages. During the critique stage researchers reflexively examine their decisions made in all other stages including why specific research questions are asked or ignored, why a given population is being studied, why certain groups are included or excluded, why the research question was operationalized into a specific study design, why a given outcome was selected for optimization, and how a model is evaluated. The critique stage asks researchers to examine their own worldview and the lens through which they conduct research. At the critique stage, the broader collective structural incentives and one's position within that society (eg, "interest and funding"[40]) help examine the other stages, design, data, and algorithm. While the critique stage is introduced last here, it is important to note that the setup calls for researchers to critique their work early and often throughout all other stages in an iterative manner.

AI4People has synthesized existing ethical ML frameworks to guide ML development and adoption.[42] The ethics axis (Figure 2) adopts a framework developed by AI4People, which combines the 4 traditional bioethics principles (beneficence, nonmaleficence, justice, and autonomy) with explicability (to better ensure intelligibility and accountability) to form a unified approach for ethical ML

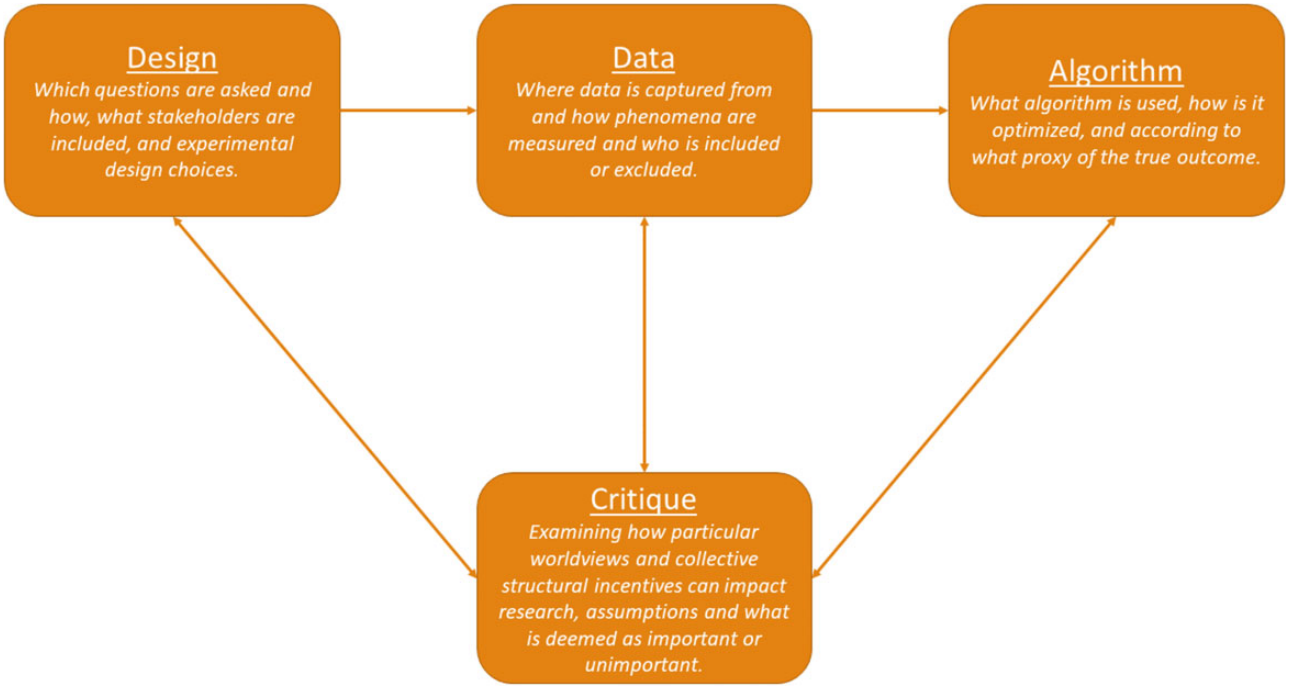

Figure 1. Proposed stages of the ML development process. Design, data, and algorithm capture stages discussed in prior work, while the critique stage incorporates Box's Loop and illustrates the cyclic nature to inherent in development. ML: machine learning.

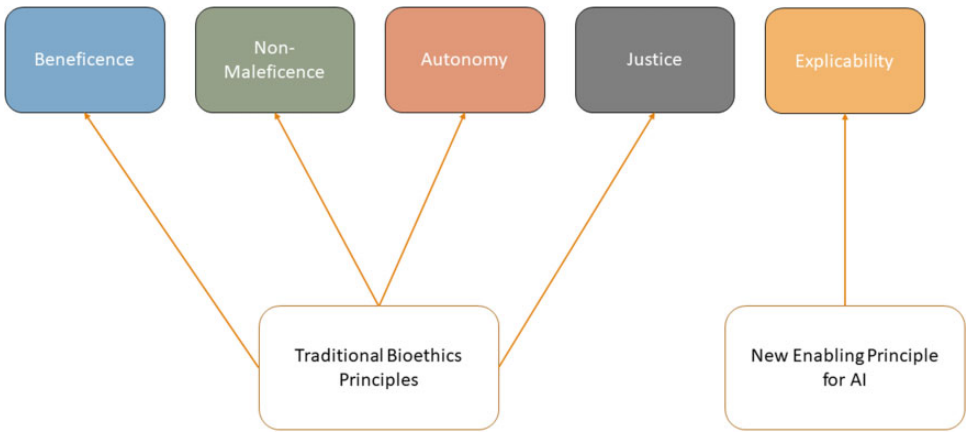

Figure 2. The ethical framework proposed by *AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations* and used to understand the complex interactions between multiple actors and clinical NLP technologies in this work. The framework focuses on the 4 traditional bioethics principles and introduces explicability to enable the other principles for application to AI. AI: artificial intelligence; NLP: natural language processing.

development and adoption.[42] We chose the AI4People framework because it combined bioethical principles used in clinical research and supported their application to ML.

In this context, beneficence pertains to designing and producing ML that benefits humanity and promotes well-being. Nonmaleficence covers aspects of good intentions gone awry and deals with preventing harms arising from ML through deliberate misuse or unpredictable behavior. Autonomy concerns human decision making and what/when decision making is ceded to ML. Justice states that ML should seek to eliminate discrimination by equally distributing the benefits and risk of healthcare resources, technologies, and datasets. Finally, explicability entails understanding ML from an epistemological perspective (how ML applications work) and as a matter of accountability (who is responsible for how ML applications work).[42] Important to this work, explicability supports the 4 bioethics principles by exposing the "technical system and the broader human process, structures, and systems around [ML]" and promoting accountability.[47]

For each article, we collected data on the study objective, NLP methods employed, the bias measure used by the authors, and the marginalized population(s) mentioned. We also extracted information relevant to the ML development process and ethical framework axes by assigning each article to one or more stage of the ML development process (if ethical considerations of that stage were mentioned in the article) and one or more ethical category (if there was implicit or explicit mention of a given principle, eg, beneficence). We then generated a matrix of analysis (Figure 3) based on assignments from that binarized decision process (applicable/not applica-

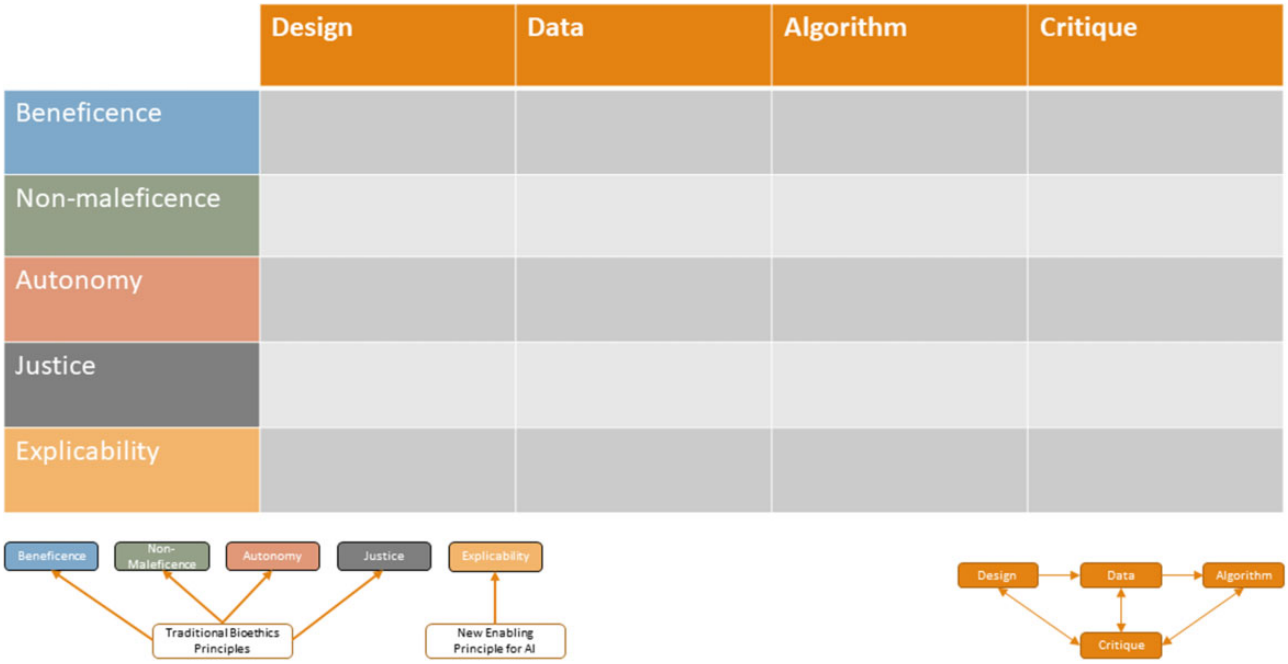

Figure 3. Articles were analyzed according to the ML development process and an ethical framework resulting in this matrix of analysis. ML: machine learning.

ble). One reviewer (OBDW) performed the initial extraction of relevant study information. A second reviewer (SSL) conducted a confirmatory assessment to evaluate consistency in data collection and assignment across studies.

RESULTS

The PRISMA ScR flow diagram (Figure 4) summarizes our study selection process. We identified and screened 1162 unique articles, conducted a full text review of 268 publications for eligibility, and selected 22 studies (Supplementary Table S1) for our main analysis.

Design

The design stage includes research question conception and experimental design choices such as the study population, stakeholder inclusion, and phenomena studied. In the design stage we identified 2 concerns: stakeholder inclusion during the design process and balancing goals with group representation.

Stakeholder inclusion is paramount to understanding important background context, study requirements, and potential pitfalls involved in research. It is especially relevant to studies concerning populations made vulnerable by systemic inequality. Three articles studied patients from LGBTQ+ communities[15–17,22] and a fourth studied geriatric patients.[18] In all cases, authors cited a lack of representation as motivation for their work.[15–18,22] No articles mentioned including patients in their research design process, though one engaged with home healthcare nurses to study attitudes and perceptions about sexual orientation and gender identity.[17] Pfohl et al[6] mentioned that technical fixes for biased models must take into account the sociopolitical contexts, and recommended including community stakeholders during the design phase.

Balancing cohort definitions and group representation begins in the design phase as researchers outline inclusion and exclusion criteria, such as requiring complete data. Though data completeness can be achieved by linking to other data sources, concerns with data rights and privacy arise.[4] As an alternative to complete data, "complete enough" data can avoid such issues, but it can also create additional biases.[4] Weber et al[4] explore how different data completeness definitions potentially bias patient-level demographic representations. The authors found that increasingly stringent data completeness standards resulted in datasets that skewed older, more female, and higher inpatient disease burden. Biased data were evaluated by comparing them to the gold standard of demographics found in the originating EHR and claims datasets.

Data

Compared to structured data, certain patient information is more accurate or only captured in clinical notes.[48,49] We identified 3 concerns in the data stage: group representation, feature representation, and biases in documentation practices. These concerns align with the notion of data justice.[50] Taylor defines 3 pillars of data justice: visibility, engagement with technology, and nondiscrimination.[50] Group and feature representation fall under the pillars of visibility and engagement with technology, as they deal with aspects if privacy and control of representation. Biased documentation practices fit under the pillar of nondiscrimination as it deals with methods and research that identifies biases.

Group representation and visibility straddles the design and data stage as improving group representation can be a motivation for carrying out research, but data are ubiquitous throughout all NLP stages so we discuss it in the data stage. Visibility was discussed in 6 papers.[4,15–17,22,51] Approaches to address poor group representation utilized keywords and structured data to identify LGBTQ+ patients.[15–17] The impact of complete data filtering on group representation was explicitly explored by one article,[4] while another article mentioned unbalanced group representation as a limitation.[51]

Data justice also raises the concern of accurate representation. Chen et al[18] sought to accurately represent older adults by identifying geriatric syndromes with a deep learning model using clinical

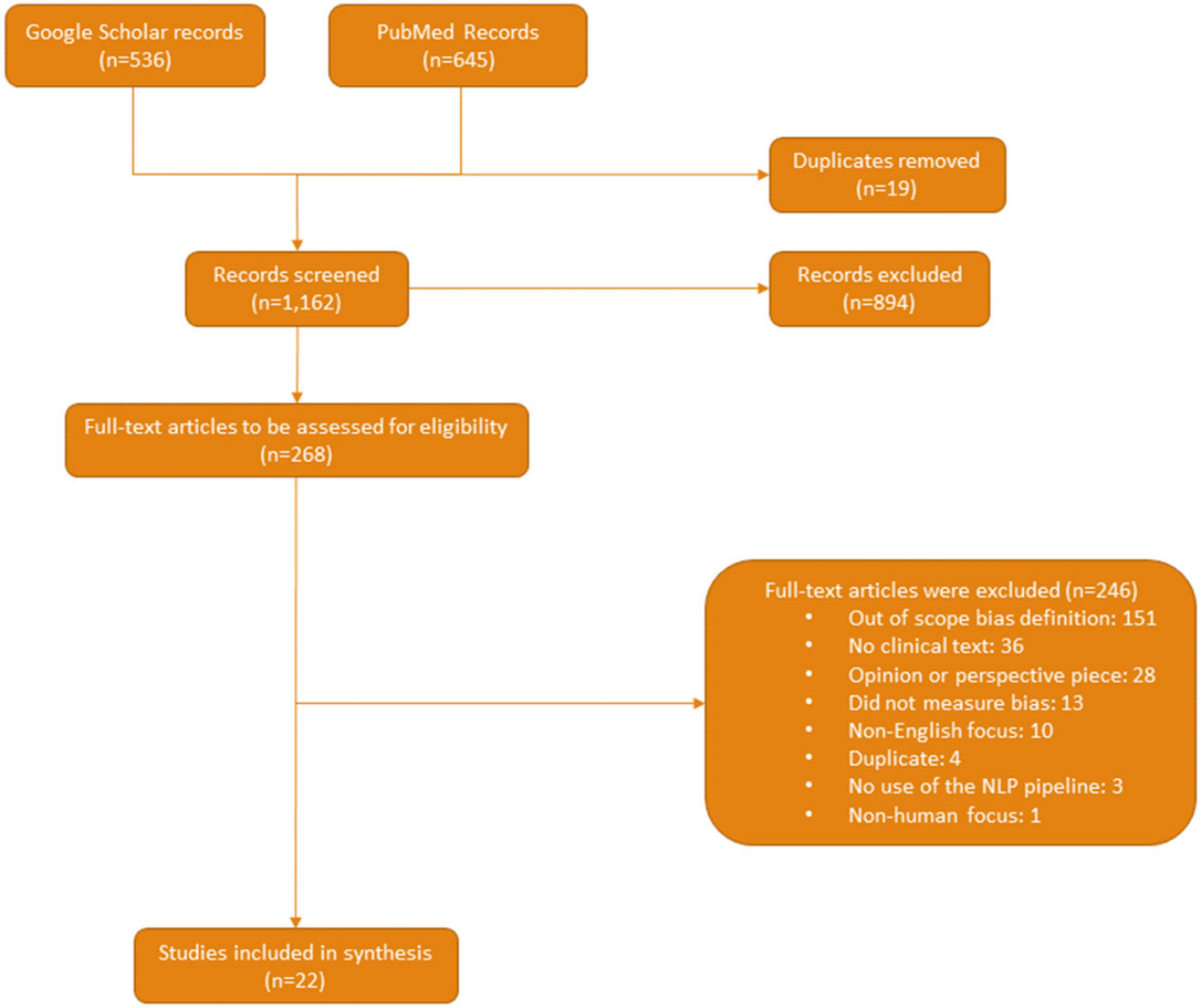

Figure 4. A flowchart of the article screening process in accordance with PRISMA guidelines.

notes and structured data. The proposed model leveraged contextual information in the document to outperform baseline models.

Clinical NLP was used to explore biased documentation practices and differences in healthcare delivery through a variety of methods such as latent Dirichlet allocation (LDA)[2,13] and differences in *n*-gram level or specific topic distributions.[13,16,17,22,52–55] LDA, an unsupervised, statistical approach to discover topics from a collection of text, was used to explore group differences in psychiatric and intensive care unit (ICU) notes[2] as well as differences between staff and residents in asthma-related discussion.[13] Sohn et al[13] found that residents were less likely to enter the diagnosis in the EHR and patients of residents had poorer outcomes compared to staff. Overall, multiple works leveraged clinical NLP to explore differences in treatment related to data justice.[13,14,52–55]

Biased documentation practices were also partially due to a lack of clinician familiarity with LGBTQ+ communities leading to omission, inaccurate, and harmful documentation.[15–17,22,55] One article used semistructured interviews to understand attitudes and perceptions about documenting LGBTQ+-specific health concerns.[17] Though sexual health discussions did not occur as often for nonheterosexual patients as heterosexual patients, building rapport somewhat alleviates this disparity.[22,55] Only 2 works (a published article[22] and dissertation[17] both by the same author) discussed stakeholder interviews focusing on providers. Other community stakeholders were not included as suggested previously.[6]

Algorithm

Algorithms can be biased for multiple reasons, including biased data[31] and choices in experimental design[10,40] (eg, choosing to model a proxy that reflects ingrained inequities or removing a racial group from analysis because of insufficient representation), as well as model selection[56] (eg, backfilling missing gender information for patients with a model that perpetuates the harmful idea of binarized gender). A majority of the work identified under the algorithm phase focused on reducing[3,5,6] or measuring[2,3,5,6,51,57] bias in applications using NLP and addressed the ethical principles of explicability and justice. We did not analyze algorithm explicability unless the presence or absence of explicability resulted in ethical considerations or explicability was a motivating factor for algorithm selection. The reason for this is that while explicability can support

Table 2. Different measures for biased models discussed throughout the work identified in this scoping review

Bias measure	Description	Relevant article(s)
Parity gap	Positive prediction differences between 2 groups	Zhang et al
Recall gap	Recall difference between 2 groups	Zhang et al
Specificity gap	Specificity difference between 2 groups	Zhang et al
AUC gap	AUC difference between 2 groups	Tsui et al
Zero-one loss gap	Zero-one loss difference between 2 groups	Chen et al
Sentence log probability gap	Difference in a language model's sentence log probability when swapping out demographic information (eg, discussion of race)	Zhang et al
Rank-turbulence divergence	Ranks occurrences of *n*-grams between 2 groups and takes into account how often rankings change	Minot et al
Conditional prediction parity	Fairness criteria that assess conditional independence between a model outcome and a demographics class. Encompasses notions of the parity gap.	Pfohl et al
Calibration fairness criteria	Measures model calibration across groups.	Pfohl et al
Cross-group ranking measures	Variation on AUC that measures how often positive instances in 1 group are ranked above negative instances in another	Pfohl et al
Sensitive attribute recovery	Measures how well a sensitive attribute (eg, gender) can be recovered	Minot et al
Demographic association with outcome	Significant association between patient demographics and model outcome using regression parameters	Wissel et al
Gold-standard bias comparison	Compare group representation to previous standard's representation	Weber et al, Polling et al

Note: This does not include measures of bias for data or healthcare delivery.

bioethical principles, explicability's effect on bias was not measured in any of the papers analyzed here.

There are many methods to measure bias and no one approach seems to be the best for all contexts.[6,32,58] We identified 8 studies that measure bias in clinical NLP models using 13 different measures (Table 2). In most cases, articles did not overlap with one another in how bias was measured.

Gap scores,[3] zero-one loss,[2] and outcome association[51] measured model outcome differences between groups. Gap scores were produced in phenotyping and mortality prediction tasks by subtracting performance metrics between 2 groups, focusing on recall gap.[3] Chen et al[2] focused on zero-one loss in psychiatric readmission and ICU mortality prediction. Outcome association was used in a single article, where the authors trained a model to identify candidates for epilepsy surgery and measured bias through the model's outputs and patient demographics using univariate and multivariate linear regressions.[51] Model decisions were not found to be significantly associated with patient demographics.[51]

Two articles used the pretrained language model, BERT,[59] which can be trained to perform new tasks in addition to the pretraining task. Zhang et al[3] found significant differences in sentence probabilities discussing various clinical categories when switching out stereotypically masculine and feminine pronouns. Minot et al[5] measured bias using the proxy of how well a trained BERT model could predict a patient's gender from the clinical notes.

Two articles used a gold-standard dataset or approach with an acceptable or normalized amount of bias,[4,60] though what constitutes a gold-standard in this scenario depends on the setting. Whereas one article measured against claims data as the most complete kind of data available,[4] another measured a model's results against a manually extracted gold-standard.[60]

One article focused on understanding bias through the lens of confounding.[61] Lynch et al[61] compared the impacts of confounding when measuring smoking status through ICD-9 codes or using information extracted from clinical notes. The authors found that when extracting smoking status as a confounder for an exposer–outcome relationship, NLP-based methods resulted in better ability to control for confounding than using ICD-9 codes for smoking status. Understanding how data sources effect confounding supports explicability.

Bias mitigation relied on methods-based approaches that directly reduced bias through either adapting the training methods or the training data. Zhang et al[3] attempted to reduce bias during the language pretraining phase. Minot et al[5] were slightly more successful in reducing their measure of bias by removing highly gendered phrases during training.

Minot et al[5] was unique in that the authors both measured bias and developed an interpretable method to mitigate bias. In particular, words which were identified as biased towards one binarized gender or another could be selectively removed during the training step to reduce bias while balancing performance on the downstream task. Their bias measure did not evaluate the differences in performance on a downstream task, but rather how well the model could recover a patient's gender from clinical notes.[5] Pfohl et al[6] also explored how different levels of bias mitigation could impact downstream task performance across a variety of measures.

As a sidenote, 2 studies cited model interpretability as a motivation to reduce bias, but neither of these studies measured bias and were not included.[15,62]

Critique

The critique stage concerns all other stages in the ML development process, as it requires researchers to examine how their positionality (ie, degrees of privilege through factors of race, class, educational attainment, income, ability, gender, and citizenship, among others[63]) and broader, collective structural incentives might have affected choices made during research. Examining one's positionality and reflection on social and structural factors are especially pertinent when studying bias. In light of this, we provide an overview of the

main critique-related concerns and topics within each stage. The main topics identified in the design stage related to motivations for research and caution for well-intentioned research that may cause harm. The data stage was associated with justification for different measures of biased data and why certain biases in datasets were addressed or not. Finally, in the algorithm stage, justification for how to measure bias and interpretability were the main topics.

Many studies included in this review were directly motivated by health disparities,[2,3,13,15–18,51,53–55] understanding data completeness,[4] or improving information extraction with automation.[14,60] Three papers discussed a lack of representation of LGBTQ+ patients, cited a dearth of applicable codes for accurate representation in structured data, and withholding information due to fear of discrimination and stigmatization.[15–17]

Sociopolitical and historical factors contribute to bias in datasets[31] and authors often used this context to explain bias. One article dismissed a limitation of unbalanced race distribution in their datasets because their race distribution matched that of US census data,[51] potentially perpetuating model bias for marginalized populations. Another article explained the differences in clinical note disease topics as representative of the medical literature.[2] One resource explored race as a proxy for other information such as mistrust, and introduced 3 measures of mistrust to characterize disparities in end-of-life care.[64] The implicit bias of clinicians can also be explored through clinical NLP and can be characterized through sentiment analysis.[65] Finally, it was found that biased language associations can be explained by differences in the disease-demographic co-occurrence statistics of training corpora.[3] An important aspect of bias in data that was not explored is the bias introduced in the selection of characteristics and dimensions of study. We recognize that "[d]ata are created and shaped by the assumptive determinations of their makers to collect some data and not others, to interrogate some objects over others and to investigate some variable relationships over others."[66] For example, investigating categories used to capture gender in clinical notes and impact of such choices are important questions in the study of biases this might perpetuate, and similar for race and ethnicity.[67–69] Our proposed critique phase of ML design creates space for such inquiry.

There are many methods for identifying bias and a lack of consensus on which is best.[6,32,58] Five studies identified biases in corpora using word and topic distributions,[2,13,14,17,22] while 2 papers examined *n*-gram level differences as a baseline analysis.[13,17] Two articles applied LDA to learn topics in corpora[2,13] and 2 studies used keyword searches or NLP models to identify specific topic differences.[13,14]

Papers which identified biased models used multiple measurements with varied motivations. One article justified using recall gap to obtain fewer false negatives for diagnostic tools and motivated their approach to evaluate language modeling bias through sentence probability by citing the approach as state-of-the-art.[3] Two articles also compared new approaches to a gold-standard approach to measure bias[4,60] and one study measured differences in group fairness using performance metrics originally motivated as clinically relevant.[57] Only one article explicitly explored the outcomes and downstream impacts of multiple fairness measures.[6] Three biased-model identification articles did not justify their measures.[2,5,51] It is important to note that measuring bias against gold-standards from EHR or claims data, data which were not originally intended to support secondary use, come with the assumption that these data and their potential biases are acceptable. This practice can perpetuate biases present in the gold-standard.

DISCUSSION

Clinical NLP can be used to study bias in applications reliant on clinical text data and explore biases in the healthcare setting. In this way, clinical NLP reflects our own biases and serves as a basis for healthcare delivery and policy changes. However, clinical NLP itself and research that leverages clinical NLP are shaped, to varying degrees, by the same forces that shape the bias we hope to study, and even work that identifies/ameliorates biases has the potential for harm. We identified several themes in each phase of the ML development process and group them into 3 main areas: (1) ambiguity when selecting metrics that interrogate and promote fairness in models; (2) opportunities and risks of active research into ML and ethics (eg, identifying sensitive patient attributes); and (3) best practices in reconciling individual autonomy, leveraging patient data, and inferring and manipulating sensitive information of subgroups.

Our review of the Design phase identified articles that explicitly focused on increasing the visibility of marginalized populations in research through clinical NLP but none of these articles discussed the potential risks for harm due to increased visibility.[15–18] Focusing on the Data phase of projects revealed that NLP can be helpful to identify and better characterize marginalized populations,[15–18] while also offering a tool to shed light on biased data generation practices.[2,13,16,17] Examination of the Algorithm phase of studies noted a plethora of methods to measure bias[2,3,6,51] and demonstrated that mitigation can be difficult.[3,5] Finally, the Critique phase illustrated that not all articles explicitly motivated their bias and explicability measures and they often presented different justifications for biased data or algorithms.[2,4,13,14,17,60]

There are multiple bias metrics that align with bioethical principles to varying degrees and for a given application some may identify bias while others do not. However, similar to the potential for clinical NLP to address and/or contribute to biases, bias metrics risk prioritizing certain ideas of what constitutes bias or harm from the perspectives of researchers and institutions that may be incongruent with those who are experiencing the outcomes of biases.[70] Similar to Metcalf et al, we also recommend researchers develop and motivate bias metrics with not only clinicians in mind, but alongside community expertise by those who are most affected by the outcomes of biases in informatics. Finally, we recommend researchers remain aware that fairness metrics, in their attempt to quantify nuanced and complex interactions, are not the gold standard to follow and reflect on, but rather a trigger for reflecting on the interactions themselves.

We also found that clinical text may also be leveraged to make visible those who were previously invisible within structured data. While improving representation can lead to more diverse research and quality care metrics, it may also expose patients to harm and discrimination. Understanding and addressing bias is not always a straight forward endeavor, and well-intentioned work may also give way to further ethical considerations. As an example, identifying gender and sexual minority patients within the EHR may violate autonomy by going against choices to withhold information due to fear of discrimination.[15–17] Moreover, once these models are in place and providing information that patients chose to withhold, they can enable harm against these patients whether it be accidental or intentional. When leveraging clinical NLP to address biases in representation, researchers should consider how such technology might be abused or misused once research is concluded and a system is in place.

Conceptions of autonomy also need to be reframed and informed by communities historically excluded from decision making regarding the use of their data. This concern is supported by the result that no articles discussed community stakeholder inclusion, besides that of providers, in their methods and only one article made the suggestion to include community stakeholders.[6] This lack of community stakeholder engagement raises concerns about who is not included in driving research into biases that ultimately impact health outcomes for patients. Individual communities are well equipped to define group status and identify potential unintended consequences of clinical NLP leveraging their data. Indigenous data sovereignty provides an example for conceptualizing community autonomy in the age of big data.[66,71]

All of these concerns point to limitations of bioethics principles for addressing group level harms that may emerge from biases in clinical NLP. Much like the existing bioethics principle we describe in this work, we note that existing ethics governing structures, like institutional review boards, also lack guidance for addressing these potential group-level harms. Thus, while the papers reviewed, and in fact most papers that use NLP and ML in healthcare, are approved by their institutional review boards, they lack critical considerations for groups, power dynamics, and systemic structures. As scholars have argued, ethical frameworks that address issues of power, recognition of groups, and structural injustices are needed to mitigate the potential for further exacerbating and reproducing inequities through technology.[37,71–74] Benjamin addresses this in the context of consent and suggests "informed refusal" as a method "to construct more reciprocal relationships between institutions and individuals"[73] for groups to address concerns over the mining of sensitive patient information such as sex and gender through clinical NLP.[15,16] This concept of refusal originated from Simpson in describing how consent was weaponized to dispossess Indigenous peoples throughout North America and Australia, and how "Refusal' rather than recognition is an option for producing and maintaining alternative structures of thought, politics and traditions away from and in critical relationship to states."[74] A current example of refusal, and specifically informed refusal, is supporting Indigenous led efforts to empower and train the next generation of Indigenous data scientists and geneticists, such as Indigidata[75] and the Summer internship for INdigenous peoples in Genomics Consortium.[76] These kinds of efforts can help community members support their community expertise with technical expertise to truly be empowered to guide, modify, and refuse research concerning their communities. Furthermore, Tsosie et al[71] argue that a bias towards prioritizing the individual as primary in bioethics is "culturally incongruent with Indigenous communitarian ethics," suggesting yet another way to reconceptualize how clinical NLP research could be conducted in consideration for its potential impact on specific communities.

The findings of this scoping review should be considered in light of several limitations. First, due to our search strategy, studies that used free-text may have been missed during the screening process if the data source was not explicitly mentioned in the title or abstract. Second, we acknowledge that some authors could have addressed our concerns and were unable to write this into the report for various reasons (audience, journal content policies, word count, etc). Either way, we cannot analyze content that did not make it into the report regardless of reason. Third, our definition of bias may differ from other authors. Lastly, while bioethical principles serve as the basis for oversight of biomedical research, we have identified limitations in applying these principles to guide research into fair clinical NLP and in creating our own reviewing methodology. In particular, we encountered a lack of guidance for evaluating how technologies interact with other technologies and assessing how sociocultural forces contribute to discrimination and other harmful outcomes.[37] This raises challenges in the consideration of group harm due to the particular emphasis of bioethics on individual autonomy.[71] To the best of our ability, we have incorporated these ideas into our analysis, but future work is need to explore other ethical frameworks.

CONCLUSIONS

Clinical NLP is a rapidly advancing field, and assessing current approaches against ethical considerations can help the discipline use clinical NLP to explore both healthcare biases and equitable NLP applications. This scoping review mapped how recent works have both studied bias in clinical NLP and used the tools of clinical NLP to study bias in healthcare delivery. Leveraging a bioethics framework and clinical ML development process, we identified challenges and opportunities in studying the intersection of clinical NLP and bias. We also recognize the limits of such frameworks for addressing potential risks of bias for groups and communities. As such, new ethics frameworks that empower communities and recognize structural injustices will be essential to intervene on the potential for clinical NLP to further entrench inequities in clinical practice.

FUNDING

This work was supported by grants from the National Library of Medicine (OBDW, HRN: T15LM007079) and National Institute of General Medical Sciences (NE: R01GM114355). The study funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. The content is solely the responsibility of the authors and does not necessarily represent the official views of a funder.

AUTHOR CONTRIBUTIONS

All authors had full access to the data in the study, and take responsibility for the integrity of the data and the accuracy of the analysis, and have approved the final manuscript. Study concept and design: OBDW and NE. Data analysis: OBDW, HRN, and SSL. Drafting and critical revision of the manuscript: all authors. Supervision: NE.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *JAMIA Open* online.

ACKNOWLEDGMENTS

We would like to thank Adrienne Pichon for her helpful feedback.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

The data underlying this article are readily available.

REFERENCES

1. Gibney E. The battle for ethical AI at the world's biggest machine-learning conference. *Nature* 2020; 577 (7792): 609–10.
2. Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? *AMA J Ethics* 2019; 21: 167–79.
3. Zhang H, Lu AX, Abdalla M, *et al*. Hurtful words: quantifying biases in clinical contextual word embeddings. In: Proceedings of the ACM Conference on Health, Inference, and Learning. New York, NY: Association for Computing Machinery; 2020: 110–20.
4. Weber GM, Adams WG, Bernstam EV, *et al*. Biases introduced by filtering electronic health records for patients with "complete data". *J Am Med Inform Assoc* 2017; 24 (6): 1134–41.
5. Minot JR, Cheney N, Maier M, *et al*. Interpretable bias mitigation for textual data: reducing gender bias in patient notes while maintaining classification performance [published online ahead of print March 9, 2021]. *arXiv:210305841 [cs, stat]*. http://arxiv.org/abs/2103.05841. Accessed July 18, 2021.
6. Pfohl SR, Foryciarz A, Shah NH. An empirical characterization of fair machine learning for clinical risk prediction. *J Biomed Inform* 2021; 113: 103621.
7. Purnell TS, Calhoun EA, Golden SH, *et al*. Achieving health equity: closing the gaps in health care disparities, interventions, and research. *Health Aff* 2016; 35 (8): 1410–5.
8. Hoppe TA, Litovitz A, Willis KA, *et al*. Topic choice contributes to the lower rate of NIH awards to African-American/black scientists. *Sci Adv* 2019; 5 (10): eaaw7238.
9. Oh SS, Galanter J, Thakur N, *et al*. Diversity in clinical and biomedical research: a promise yet to be fulfilled. *PLOS Med* 2015; 12 (12): e1001918.
10. Obermeyer Z, Powers B, Vogeli C, *et al*. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019; 366 (6464): 447–53.
11. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight – reconsidering the use of race correction in clinical algorithms. *N Engl J Med* 2020; 383 (9): 874–82.
12. Goldberg Y. Neural network methods for natural language processing. In: Synthesis Lectures on Human Language Technologies, Vol. 10. 2017: 1–309. doi:10.2200/S00762ED1V01Y201703HLT037.
13. Sohn S, Wi C-I, Juhn YJ, *et al*. Analysis of clinical variations in asthma care documented in electronic health records between staff and resident physicians. *Stud Health Technol Inform* 2017; 245: 1170–4.
14. Skaljic M, Patel IH, Pellegrini AM, *et al*. Prevalence of financial considerations documented in primary care encounters as identified by natural language processing methods. *JAMA Netw Open* 2019; 2 (8): e1910399e1910399.
15. Guo Y, He X, Lyu T, *et al*. Developing and validating a computable phenotype for the identification of transgender and gender nonconforming individuals and subgroups. *AMIA Annu Symp Proc* 2021;2020:514–23.
16. Ehrenfeld JM, Gottlieb KG, Beach LB, *et al*. Development of a natural language processing algorithm to identify and evaluate transgender patients in electronic health record systems. *Ethn Dis* 2019; 29 (Suppl 2): 441–50.
17. Bjarnadottir RI. *Assessment and documentation of sexual orientation and gender identity in home healthcare* [dissertation]. Columbia University; 2016. doi:10.7916/D8ZW1M3V.
18. Chen T, Dredze M, Weiner JP, *et al*. Identifying vulnerable older adult populations by contextualizing geriatric syndrome information in clinical notes of electronic health records. *J Am Med Inform Assoc* 2019; 26 (8–9): 787–95.
19. Flynn RWV, Macdonald TM, Schembri N, *et al*. Automated data capture from free-text radiology reports to enhance accuracy of hospital inpatient stroke codes. *Pharmacoepidemiol Drug Saf* 2010; 19 (8): 843–7.
20. Yang H, Spasic I, Keane JA, *et al*. A text mining approach to the prediction of disease status from clinical discharge summaries. *J Am Med Inform Assoc* 2009; 16 (4): 596–600.
21. Friedman C, Alderson PO, Austin JH, *et al*. A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc* 1994; 1 (2): 161–74.
22. Bjarnadottir RI, Bockting W, Yoon S, *et al*. Nurse documentation of sexual orientation and gender identity in home healthcare: a text mining study. *Comput Inform Nurs* 2019; 37 (4): 213–21.
23. Ou Y, Patrick J. Automatic structured reporting from narrative cancer pathology reports. *Electron J Health Inform* 2014; 8: e20.
24. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet* 2012; 13 (6): 395–405.
25. Ye C, Fu T, Hao S, *et al*. Prediction of incident hypertension within the next year: prospective study using statewide electronic health records and machine learning. *J Med Internet Res* 2018; 20 (1): e22.
26. Torii M, Fan J, Yang W, *et al*. Risk factor detection for heart disease by applying text analytics in electronic medical records. *J Biomed Inform* 2015; 58: S164–70.
27. Miotto R, Li L, Kidd BA, *et al*. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci Rep* 2016; 6: 26094.
28. Jensen K, Soguero-Ruiz C, Oyvind Mikalsen K, *et al*. Analysis of free text in electronic health records for identification of cancer patient trajectories. *Sci Rep* 2017; 7: 46226–12.
29. Imler TD, Morea J, Imperiale TF. Clinical decision support with natural language processing facilitates determination of colonoscopy surveillance intervals. *Clin Gastroenterol Hepatol* 2014; 12 (7): 1130–6.
30. Ben Abacha A, Zweigenbaum P. MEANS: a medical question-answering system combining NLP techniques and semantic Web technologies. *Inform Process Manag* 2015; 51 (5): 570–94.
31. Mehrabi N, Morstatter F, Saxena N, *et al*. A survey on bias and fairness in machine learning [published online ahead of print September 17, 2019]. *arXiv:190809635 [cs]*. http://arxiv.org/abs/1908.09635. Accessed December 18, 2020.
32. Corbett-Davies S, Goel S. The measure and mismeasure of fairness: a critical review of fair machine learning [published online ahead of print August 14, 2018]. *arXiv:180800023 [cs]*. http://arxiv.org/abs/1808.00023. Accessed August 16, 2020.
33. Pedreshi D, Ruggieri S, Turini F. *Discrimination-aware data mining. In: proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining*; 2008: 560–8; New York, NY: Association for Computing Machinery.
34. Hovy D, Prabhumoye S. Five sources of bias in natural language processing. *Lang Linguist Compass* 2021; 15 (8): e12432.
35. Floridi L. *The Ethics of Informaiton*. Oxford: Oxford University Press; 2013.
36. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *International J Soc Res Methodol* 2005; 8 (1): 19–32.
37. Hoffmann AL. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Inform Commun Soc* 2019; 22 (7): 900–15.
38. Tricco AC, Lillie E, Zarin W, *et al*. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med* 2018; 169 (7): 467–73.
39. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Cambridge, MA: MIT Press; 2016.
40. Chen IY, Pierson E, Rose S, *et al*. Ethical machine learning in health care [published online ahead of print September 23, 2020]. *arXiv:200910576 [cs]*. http://arxiv.org/abs/2009.10576. Accessed September 30, 2020).
41. Blei DM. Build, compute, critique, repeat: data analysis with latent variable models. *Annu Rev Stat Appl* 2014; 1 (1): 203–32.
42. Floridi L, Cowls J, Beltrametti M, *et al*. AI4People – an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds Mach* 2018; 28 (4): 689–707.
43. Mullainathan S, Obermeyer Z. On the inequity of predicting a while hoping for B. *AEA Papers Proc* 2021; 111: 37–42.
44. Alsentzer E, Murphy J, Boag W, *et al*. Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. Minneapolis, MN: Association for Computational Linguistics; 2019: 72–8.
45. Dligach D, Afshar M, Miller T. Toward a clinical text encoder: pretraining for clinical natural language processing with applications to substance misuse. *J Am Med Inform Assoc* 2019; 26 (11): 1272–8.

46. Pakhomov S, Shah N, Hanson P, *et al.* Automatic quality of life prediction using electronic medical records. *AMIA Annu Symp Proc* 2008; 2008: 545–9.
47. Cobbe J, Singh J. Reviewable automated decision-making. *Comp Law Security Rev* 2020; 39: 105475.
48. Walsh C, Elhadad N. Modeling clinical context: rediscovering the social history and evaluating language from the clinic to the wards. *AMIA Jt Summits Transl Sci Proc* 2014; 2014: 224–31.
49. Klinger EV, Carlini SV, Gonzalez I, *et al.* Accuracy of race, ethnicity, and language preference in an electronic health record. *J Gen Intern Med* 2015; 30 (6): 719–23.
50. Taylor L. What is data justice? The case for connecting digital rights and freedoms globally. *Big Data Soc* 2017; 4 (2): 2053951717736335.
51. Wissel BD, Greiner HM, Glauser TA, *et al.* Investigation of bias in an epilepsy machine learning algorithm trained on physician notes. *Epilepsia* 2019; 60 (9): e93–8.
52. Werbeloff N, Hilge Thygesen J, Hayes JF, *et al.* Childhood sexual abuse in patients with severe mental illness: demographic, clinical and functional correlates. *Acta Psychiatr Scand* 2021; 143 (6): 495–502.
53. Irving J, Colling C, Shetty H, *et al.* Gender differences in clinical presentation and illicit substance use during first episode psychosis: a natural language processing, electronic case register study. *BMJ Open* 2021; 11 (4): e042949.
54. Wellesley Wesley E, Patel I, Kadra-Scalzo G, *et al.* Gender disparities in clozapine prescription in a cohort of treatment-resistant schizophrenia in the South London and Maudsley case register. *Schizophr Res* 2021; 232: 68–76.
55. Lynch KE, Viernes B, Schliep KC, *et al.* Variation in sexual orientation documentation in a national electronic health record system. *LGBT Health* 2021; 8 (3): 201–8.
56. Keyes O. The misgendering machines: trans/HCI implications of automatic gender recognition. *Proc ACM Hum–Comput Interact* 2018; 2: article 88:
57. Tsui FR, Shi L, Ruiz V, *et al.* Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts. *JAMIA Open* 2021; 4 (1): ooab011.
58. Gonen H, Goldberg Y. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove rhem. 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). Minneapolis, MN: Association for Computational Linguistics; 2019. 609–14.
59. Devlin J, Chang M-W, Lee K, *et al.* 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). Minneapolis, MN: Association for Computational Linguistics; 2019: 4171–86.
60. Polling C, Tulloch A, Banerjee S, *et al.* Using routine clinical and administrative data to produce a dataset of attendances at emergency departments following self-harm. *BMC Emerg Med* 2015; 15: 15.
61. Lynch KE, Whitcomb BW, DuVall SL. How confounder strength can affect allocation of resources in electronic health records. *Perspect Health Inf Manag* 2018; 15: 1d.
62. Gehrmann S, Dernoncourt F, Li Y, *et al.* Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PLoS One* 2018; 13 (2): e0192360.
63. Duarte ME. *Network Sovereignty: Building the Internet across Indian Country*. Seattle, WA: University of Washington Press; 2017.
64. Boag WWG. Quantifying racial disparities in end-of-life care; 2018. https://dspace.mit.edu/handle/1721.1/118063. Accessed May 17, 2021.
65. Weissman GE, Ungar LH, Harhay MO, *et al.* Construct validity of six sentiment analysis methods in the text of encounter notes of patients with critical illness. *J Biomed Inform* 2019; 89: 114–21.
66. Walter M, Kukutai T, Carroll SR, *et al. Indigenous Data Sovereignty and Policy*. London: Routledge; 2020.
67. Boland MR, Elhadad N, Pratt W. Informatics for sex- and gender-related health: understanding the problems, developing new methods, and designing new solutions. *J Am Med Inform Assoc* 2022; 29 (2): 225–9.
68. Tatonetti NP, Elhadad N. Fine-scale genetic ancestry as a potential new tool for precision medicine. *Nat Med* 2021; 27 (7): 1152–3.
69. Boehmer U, Kressin NR, Berlowitz DR, *et al.* Self-reported vs administrative race/ethnicity data and study results. *Am J Public Health* 2002; 92 (9): 1471–2.
70. Metcalf J, Moss E, Watkins EA, *et al.* Algorithmic impact assessments and accountability: the co-construction of impacts. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. New York, NY: Association for Computing Machinery; 2021: 735–46.
71. Tsosie KS, Claw KG, Garrison NA. Considering "Respect for Sovereignty" beyond the Belmont report and the common rule: ethical and legal implications for American Indian and Alaska native peoples. *Am J Bioeth* 2021; 21 (10): 27–30.
72. TallBear K. *Native American DNA: Tribal Belonging and the False Promise of Genetic Science*. Minneapolis, MN: University of Minnesota Press; 2013.
73. Benjamin R. Informed refusal: toward a justice-based bioethics. *Sci Technol Hum Values* 2016; 41 (6): 967–90.
74. Simpson A. The ruse of consent and the anatomy of 'refusal': cases from indigenous North America and Australia. *Postcolon Stud* 2017; 20 (1): 18–33.
75. IndigiData—Indigenous data science education. IndigiData. https://indigidata.nativebio.org/. Accessed May 2, 2022.
76. Malhi RS, Bader A. Engaging native Americans in genomics research. *Am Anthropol* 2015; 117 (4): 743–4.