Algorithmic Fairness in Clinical Natural Language Processing: Challenges and Opportunities

Daniel Anadria[1,2,*], Anastasia Giachanou[1], Jacqueline Kernahan[3,2], Roel Dobbe[3] and Daniel Oberski[1]

[1]Utrecht University
[2]University Medical Center Utrecht
[3]Delft University of Technology

Abstract
The surge in research and development of clinical natural language processing (NLP) has prompted inquiries into the algorithmic fairness of the proposed and deployed technical solutions. In spite of the proliferation of research, limited work has synthesized reflected on the state of *algorithmic fairness in clinical NLP*. In this short paper, we summarize the findings of our scoping review of literature and present challenges and opportunities in the domain. We identify challenges and opportunities related to studying and measuring protected groups, selecting appropriate methodology, data sharing and privacy, as well as generalizability. The goal of this article is to start a discussion and raise awareness about the gaps encountered within algorithmic fairness in clinical NLP and pave the way for future research.

Keywords
clinical natural language processing, algorithmic fairness, research gaps, NLP in healthcare

1. Introduction

Clinical text, i.e. clinician-generated writing about patients, such as that found in electronic health records and clinical notes, is a rich source of unstructured patient data. NLP pipelines can leverage latent signals in clinical text to extract information for decision support tools used in patient care and clinical research. Recent advancements in large language models have paved the way for novel clinical applications of natural language generation [1, 2, 3, 4]. Many studies have demonstrated the effectiveness of NLP on tasks such as pathology detection and risk prediction [5, 6, 7, 8], extraction of social determinants of health from electronic health records [9, 10, 11, 12], and generation of patient discharge summaries [2, 13, 14].

NLP algorithms recognize and leverage various patterns that are encoded within clinical text. While their ability to discover correlations in the data structure enables statistical modeling

EWAF'24: European Workshop on Algorithmic Fairness, July 01–03, 2024, Mainz, Germany
*Corresponding author.
d.anadria@uu.nl (D. Anadria); a.giachanou@uu.nl (A. Giachanou); j.a.kernahan@tudelft.nl (J. Kernahan); r.i.j.dobbe@tudelft.nl (R. Dobbe); d.l.oberski@uu.nl (D. Oberski)
https://danadria.com/ (D. Anadria); https://www.uu.nl/staff/AGiachanou (A. Giachanou); https://www.uu.nl/staff/DLOberski (D. Oberski)
0009-0008-2247-4887 (D. Anadria); 0000-0002-7601-8667 (A. Giachanou); 0009-0003-5840-2299 (J. Kernahan); 0000-0003-4633-7023 (R. Dobbe); 0000-0001-7467-2297 (D. Oberski)

of natural language, when learned associations are spurious or their normative implications are considered illegitimate within the current context, this same ability also limits the validity of derived prediction and inference. In addition to the valuable medical signal, data capture noise which reflects discrepancies due to past and current social realities that had influenced the data generation process [15]. This noise can encompass various *data biases* including socio-economic health inequities and social determinants of health [16, 17, 18], differential care seeking behavior [16, 19], differences in language physicians use to describe patients [20, 21, 22], differences in treatment physicians prescribe to different population segments [23, 24, 25], as well as variability in clinical presentation of diseases [26], and adverse drug reactions [27]. A clinical NLP pipeline can be considered fair if it neither automates nor perpetuates social stigma and stereotyping of patient groups, constituting *representational harms*, nor systematically denies patient groups access to opportunities and resources, causing *allocative harms* [28].

Clinical NLP pipelines need to be developed responsibly with robust safety, validity, and fairness checks in order to ensure that the NLP solution does not automate nor amplify existing inequities leading to harm. Fortunately, just as clinical NLP pipelines can propagate existing healthcare inequities encoded in the data [29, 30], in some cases the same pipelines can be tinkered with to produce outcomes more equitable than those of the existing healthcare systems which had generated the training data [1] [31]. To this end, previous studies on algorithmic fairness in clinical NLP have proposed a variety of fairness auditing and bias mitigation frameworks [29, 30, 31, 32, 33, 34, 35, 36, 37].

In spite of the growing interest in fairness of clinical NLP tools and pipelines, there is a scarcity of evidence synthesis in this domain. To the best of our knowledge, only one review [38] focusing on fairness in clinical NLP has been published to date, largely centering on ethical considerations surrounding pipeline development. The present work aims to identify gaps in fair clinical NLP research. We conducted a systematic literature search spanning six scholarly databases (PubMed, Embase, Web of Science, Scopus, ACM Digital Library, and IEEE Xplore) and three search engines (Google Scholar, Semantic Scholar, and Scholar AI). Our query terms were related to the concepts of *NLP*, *fairness*, and *healthcare*. The search took place between 18 and 25 October 2023, and it resulted in 355 unique papers - 24 of which were deemed to be core inclusions, i.e. applied studies using NLP for a clinical task involving patient data and assessing the fairness of the NLP pipeline. The search had also identified a number of theoretical papers that have proven relevant for the identification of challenges and opportunities in the domain.

2. Challenges and opportunities

In this section, we discuss the gaps of fair clinical NLP research as identified in our review of the literature. Each gap reflect a challenge, as well as an opportunity for future research.

Protected groups. As establishing fairness of algorithmic representations and outcomes across all demographic groups might not be feasible [39], the choice of which groups the

[1]The NLP community efforts primarily focus on harm detection and mitigation at the level of data and models - usually targeting model representations of protected groups and the distribution of model-assigned outcomes. In practice, harm can also appear elsewhere in the development cycle, for instance as a result of an incorrect solution deployment [15]. A broader socio-technical lens can help explain how NLP interventions depend on upstream activities, and shape downstream activities and outcomes.

clinical NLP pipeline should be demonstrated to treat equitably is of increasing importance. We find that the groups examined in the literature are narrow in scope, with the majority of studies focusing on gender, race/ethnicity, and to a lesser extent, age. In particular, research primarily concerned US-centric protected groups. Vulnerable groups such as individuals with mental illness diagnoses [40, 41, 42], various forms of disability [43], or traditionally overlooked groups, such as individuals admitted during the weekend as opposed to on a weekday [44] remain underrepresented in the clinical fairness literature. Furthermore, the difference in the geographical and cultural context on which local demographics should be considered protected remains under-examined. The variability in how groups are conceptualized and treated is significant both within and between societies, and groups marginalized in some contexts may not be recognized as such in others. The studies have focused on the more numerous disadvantaged groups, which is in line with the utilitarian goal of maximizing the well-being of the greatest number of individuals [2] [45, 46]. However, this leaves a gap when it comes to protecting smaller-sized groups such as those at the intersection of multiple disenfranchised identities [47]. Future research should encompass groups broader than those defined by sex, race/ethnicity, and age, and explore intersectional, understudied, and biases affecting smaller-sized groups. Importantly, the choice of whom to protect should always be motivated by the local clinical and broader societal context surrounding the NLP pipeline development.

An additional challenge arises from the imperfect measurement of group membership, which ranges from being fully absent to the use of various proxies [48]. Previous studies have examined the construction of common group labels, such as race [49] and gender [50] and the associated noise. In healthcare, NLP has also been used to construct missing labels [12, 51, 52, 53]. We find that the majority of inclusions did not report how the attribute labels had been constructed in the data generation process, except for those where authors created their own labels using regular expressions or string searches. Ideally, clinical datasets would include information on social determinants of health as their inclusion has been shown to improve fairness for vulnerable groups [54]. Some clinical NLP studies rely on self-reported labels which might limit their validity in certain situations [49]. In cases where protected attribute information is fully absent, data imputation methods, such as Bayesian Improved Surname Geocoding [55, 56] can estimate group membership based on relevant correlates. Similar tools are needed for countries beyond the US. The development of robust indirect estimation methods is necessary to achieve attribute data completeness, a prerequisite for conducting fairness audits and harm mitigations.

Method selection. Fairness auditing and harm mitigation carry many researcher degrees of freedom. While inclusions take on operational definitions of bias and fairness, and in some cases propose debiasing methods, we find that the motivations behind these choices are seldom reported in the literature. Furthermore, not every computationally feasible approach might have clinical legitimacy. For instance, Minot et al. [34] performed a naive removal of the most-gendered tokens [34]. While the approach removed terms such as "he", "his", "she", and "her", it had also erased medically valuable terms such as "urinal", "prostate", "hysterectomy", "vaginal", and "osteoporosis". Such approaches might lead to a loss of valuable clinical information. At the present time, there is a lack of clarity as to under what conditions could a certain method

[2] An important limitation is that, in many pipelines, the most frequently studied demographic groups might be the only ones with available information for conducting a fairness audit.

be considered appropriate. While we observe a plethora of fairness metrics and bias mitigation methodologies within the clinical NLP literature [29, 30, 31, 32, 33, 34, 35, 36, 37], we also note that the majority of inclusions do not motivate their methodological choices. The presence of algorithmic bias should be corroborated with an understanding of its source as this can help inform the appropriate mitigation approach [57]. There is a need for openness and transparency.

Data sharing and privacy. A key challenge for clinical NLP developers is the acquisition of diverse real world datasets, especially those containing protected attribute information. Data sharing is frequently predicated on a degree of de-identification of confidential patient records [58]. Ironically, even patient de-identification using NLP has been shown to not be equally effective across patient demographics [59]. Algorithmic fairness research requires access to the very same sensitive information which healthcare institutions might prefer anonymized prior to data sharing. We call for a great inclusion of sensitive attributes in clinical datasets as this can help develop accurate and safe clinical decision support systems.

Another challenge lies in the construction of accurate outcome labels for supervised learning tasks, especially in large datasets where expert annotation becomes increasingly expensive. Privacy restrictions limit the public availability of real world datasets. The sharing of more text-rich clinical datasets would enhance the development of fair clinical NLP, but this needs to be balanced with patient privacy concerns. Synthetic data is one potential solution to address this challenge [60, 61]. Also, methodological solutions such as transfer learning and weak supervision approaches might help alleviate the problem of the missing gold standard.

Generalizability. The lack of diversity in clinical NLP datasets poses a major limitation to the literature. MIMIC [62] and MIMIC-derived datasets [63, 64, 65, 66, 67, 68] represented the majority of publicly available free text data. Our search has revealed few publicly-available English language datasets not based on MIMIC notes [37, 69, 70]. While some of the inclusions had access to non-public hospital data, in all the studies the hospitals were based in the US. This speaks to the gap in research on languages other than English, and countries other than the US. Our additional search of PhysioBank [71] has revealed that the only languages with public medical databases other than English were Spanish [72] and Brazilian Portuguese [73], each with a single clinical database. This lack of research beyond English can create a major problem for the generalizability of the developed tools and methodologies that might underperform in languages different from English. We call for more research on languages other than English, societies other than the US, and patient demographics beyond the US protected groups.

3. Conclusion

This short paper summarizes the findings of our scoping review investigating challenges and opportunities for algorithmic fairness in clinical NLP. We have identified gaps related to studying and measuring protected groups, selecting appropriate methodology, data sharing and privacy, as well as generalizability. While algorithmic fairness in clinical NLP comes with many challenges, most of these also carry inherent opportunities for future research. We hope to start a discussion within the algorithmic fairness community and direct future work towards closing the gaps.

References

[1] H. ten Berg, B. van Bakel, L. van de Wouw, K. E. Jie, A. Schipper, H. Jansen, R. D. O'Connor, B. van Ginneken, S. Kurstjens, ChatGPT and Generating a Differential Diagnosis Early in an Emergency Department Presentation, Annals of Emergency Medicine 83 (2024) 83–86. URL: https://www.sciencedirect.com/science/article/pii/S019606442300642X. doi:10.1016/j.annemergmed.2023.08.003.

[2] S. Singh, A. Djalilian, M. J. Ali, ChatGPT and Ophthalmology: Exploring Its Potential with Discharge Summaries and Operative Notes, Seminars in Ophthalmology 38 (2023) 503–507. URL: https://doi.org/10.1080/08820538.2023.2209166.

[3] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. D. Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, V. Tseng, Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models, PLOS Digital Health 2 (2023) 1–12. URL: https://doi.org/10.1371/journal.pdig.0000198.

[4] S. B. Patel, K. Lam, ChatGPT: the future of discharge summaries?, The Lancet Digital Health 5 (2023) e107–e108. URL: https://doi.org/10.1016/S2589-7500(23)00021-3.

[5] P. Mosteiro, E. Rijcken, K. Zervanou, U. Kaymak, F. Scheepers, M. Spruit, Machine Learning for Violence Risk Assessment Using Dutch Clinical Notes, Journal of Artificial Intelligence for Medical Sciences 2 (2021) 44–54. URL: https://doi.org/10.2991/jaims.d.210225.001.

[6] A. Bagheri, T. K. J. Groenhof, F. W. Asselbergs, S. Haitjema, M. L. Bots, W. B. Veldhuis, P. A. de Jong, D. L. Oberski, Automatic Prediction of Recurrence of Major Cardiovascular Events: A Text Mining Study Using Chest X-Ray Reports, Journal of Healthcare Engineering 2021 (2021) 1–11. URL: https://doi.org/10.1155/2021/6663884, publisher: Hindawi.

[7] V. Menger, F. Scheepers, M. Spruit, Comparing Deep Learning and Classical Machine Learning Approaches for Predicting Inpatient Violence Incidents from Clinical Text, Applied Sciences 8 (2018) 981. URL: https://doi.org/10.3390/app8060981.

[8] J. Irving, R. Patel, D. Oliver, C. Colling, M. Pritchard, M. Broadbent, H. Baldwin, D. Stahl, R. Stewart, P. Fusar-Poli, Using Natural Language Processing on Electronic Health Records to Enhance Detection and Prediction of Psychosis Risk, Schizophrenia Bulletin 47 (2021) 405–414. URL: https://doi.org/10.1093/schbul/sbaa126.

[9] K. S. Allen, D. R. Hood, J. Cummins, S. Kasturi, E. A. Mendonca, J. R. Vest, Natural language processing-driven state machines to extract social factors from unstructured clinical documentation, JAMIA Open 6 (2023) 1–8. URL: https://doi.org/10.1093/jamiaopen/ooad024.

[10] W. Wu, K. J. Holkeboer, T. O. Kolawole, L. Carbone, E. Mahmoudi, Natural language processing to identify social determinants of health in Alzheimer's disease and related dementia from electronic health records, Health Services Research Epub ahead of print (2023). URL: https://doi.org/10.1111/1475-6773.14210.

[11] K. Lybarger, N. J. Dobbins, R. Long, A. Singh, P. Wedgeworth, Uzuner, M. Yetisgen, Leveraging natural language processing to augment structured social determinants of health data in the electronic health record, Journal of the American Medical Informatics Association: JAMIA 30 (2023) 1389–1397. URL: https://doi.org/10.1093/jamia/ocad073.

[12] Z. Sheng, NLP System for Mining Social Determinant of Health From Clinical Notes and its Fairness Evaluations, in: 2022 IEEE 10th International Conference on Healthcare

Informatics (ICHI), 2022, pp. 478–478. URL: https://doi.org/10.1109/ICHI54592.2022.00076.
[13] T. Searle, Z. Ibrahim, J. Teo, R. J. B. Dobson, Discharge summary hospital course summarisation of in patient Electronic Health Record text with clinical concept guided deep pre-trained Transformer models, Journal of Biomedical Informatics 141 (2023) 104358. URL: https://doi.org/10.1016/j.jbi.2023.104358. doi:10.1016/j.jbi.2023.104358.
[14] G. S. Rosenberg, M. Magnéli, N. Barle, M. G. Kontakis, A. M. Müller, M. Wittauer, M. Gordon, C. Brodén, ChatGPT-4 generates orthopedic discharge documents faster than humans maintaining comparable quality: a pilot study of 6 cases, Acta Orthopaedica 95 (2024) 152–156. URL: https://doi.org/10.2340/17453674.2024.40182.
[15] H. Suresh, J. Guttag, A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle, in: Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, EAAMO '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 1–9. URL: https://doi.org/10.1145/3465416.3483305.
[16] OECD, Health for Everyone?: Social Inequalities in Health and Health Systems, Technical Report, Organisation for Economic Co-operation and Development, Paris, 2019. URL: https://doi.org/10.1787/3c8385d0-en.
[17] Y. Huang, J. Guo, Z. Chen, J. Xu, W. T. Donahoo, O. Carasquillo, H. Adloori, J. Bian, E. A. Shenkman, The impact of electronic health records (EHR) data continuity on prediction model fairness and racial-ethnic disparities, 2023. URL: https://doi.org/10.48550/arXiv.2309.01935.
[18] G. McCartney, F. Popham, R. McMaster, A. Cumbers, Defining health and health inequalities, Public Health 172 (2019) 22–30. URL: https://doi.org/10.1016/j.puhe.2019.03.023.
[19] P. M. Galdas, F. Cheater, P. Marshall, Men and health help-seeking behaviour: literature review, Journal of Advanced Nursing 49 (2005) 616–623. URL: https://doi.org/10.1111/j.1365-2648.2004.03331.x.
[20] G. Himmelstein, D. Bates, L. Zhou, Examination of Stigmatizing Language in the Electronic Health Record, JAMA network open 5 (2022) e2144967. URL: https://10.1001/jamanetworkopen.2021.44967.
[21] M. C. Beach, S. Saha, J. Park, J. Taylor, P. Drew, E. Plank, L. A. Cooper, B. Chee, Testimonial Injustice: Linguistic Bias in the Medical Records of Black Patients and Women, Journal of General Internal Medicine 36 (2021) 1708–1714. URL: https://10.1007/s11606-021-06682-z.
[22] V. Barcelona, D. Scharp, B. R. Idnay, H. Moen, D. Goffman, K. Cato, M. Topaz, A qualitative analysis of stigmatizing language in birth admission clinical notes, Nursing Inquiry 30 (2023) 1–11. URL: https://doi.org/10.1111/nin.12557.
[23] E. N. Chapman, A. Kaatz, M. Carnes, Physicians and Implicit Bias: How Doctors May Unwittingly Perpetuate Health Care Disparities, Journal of General Internal Medicine 28 (2013) 1504–1510. URL: https://doi.org/10.1007/s11606-013-2441-1.
[24] K. A. Schulman, J. A. Berlin, W. Harless, J. F. Kerner, S. Sistrunk, B. J. Gersh, R. Dubé, C. K. Taleghani, J. E. Burke, S. Williams, J. M. Eisenberg, J. J. Escarce, The effect of race and sex on physicians' recommendations for cardiac catheterization, The New England Journal of Medicine 340 (1999) 618–626. URL: https://doi.org/10.1056/NEJM199902253400806.
[25] P. Lee, M. Le Saux, R. Siegel, M. Goyal, C. Chen, Y. Ma, A. C. Meltzer, Racial and ethnic disparities in the management of acute pain in US emergency departments: Meta-analysis

and systematic review, The American Journal of Emergency Medicine 37 (2019) 1770–1777. URL: https://doi.org/10.1016/j.ajem.2019.06.014.

[26] W. Chen, S. L. Woods, D. J. Wilkie, K. A. Puntillo, Gender Differences in Symptom Experiences of Patients with Acute Coronary Syndromes, Journal of Pain and Symptom Management 30 (2005) 553–562. URL: https://doi.org/10.1016/j.jpainsymman.2005.06.004.

[27] S. H. Bots, F. Groepenhoff, A. L. M. Eikendal, C. Tannenbaum, P. A. Rochon, Z. V. Regitz, V. M. Miller, D. Day, F. W. Asselbergs, R. H. M. den, Adverse Drug Reactions to Guideline-Recommended Heart Failure Drugs in Women, JACC: Heart Failure 7 (2019) 258–266. URL: https://doi.org/10.1016/j.jchf.2019.01.009.

[28] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities, MIT Press, Cambridge, MA, USA, 2023.

[29] H. Zhang, A. X. Lu, M. Abdalla, M. McDermott, M. Ghassemi, Hurtful words: quantifying biases in clinical contextual word embeddings, in: Proceedings of the ACM Conference on Health, Inference, and Learning, CHIL '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 110–120. URL: https://10.1145/3368555.3384448.

[30] H. Adam, M. Y. Yang, K. Cato, I. Baldini, C. Senteio, L. A. Celi, J. Zeng, M. Singh, M. Ghassemi, 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, AIES '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 7–21. URL: https://doi.org/10.1145/3514094.3534203.

[31] S. Agmon, P. Gillis, E. Horvitz, K. Radinsky, Gender-sensitive word embeddings for healthcare, Journal of the American Medical Informatics Association 29 (2022) 415–423. URL: https://doi.org/10.1093/jamia/ocab279.

[32] J. Chen, I. Berlot-Attwell, X. Wang, S. Hossain, F. Rudzicz, Exploring Text Specific and Blackbox Fairness Algorithms in Multimodal Clinical NLP, in: Proceedings of the 3rd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Online, 2020, pp. 301–312. URL: https://doi.org/10.18653/v1/2020.clinicalnlp-1.33.

[33] E. Ferracane, S. Konam, Towards Fairness in Classifying Medical Conversations into SOAP Sections, 2020. URL: https://doi.org/10.48550/arXiv.2012.07749, arXiv:2012.07749 [cs].

[34] J. R. Minot, N. Cheney, M. Maier, D. C. Elbers, C. M. Danforth, P. S. Dodds, Interpretable Bias Mitigation for Textual Data: Reducing Genderization in Patient Notes While Maintaining Classification Performance, ACM Transactions on Computing for Healthcare 3 (2022) 39:1–39:41. URL: https://doi.org/10.1145/3524887.

[35] G. Sogancioglu, A. A. Salah, The effects of gender bias in word embeddings on patient phenotyping in the mental health domain, in: 2023 11th International Conference on Affective Computing + Intelligent Interaction (ACII), Cambridge, MA, USA, 2023. URL: https://doi.org/10.48550/arXiv.2212.07852.

[36] R. Poulain, M. F. Bin Tarek, R. Beheshti, Improving Fairness in AI Models on Electronic Health Records: The Case for Federated Learning Methods, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 1599–1608. URL: https://doi.org/10.1145/3593013.3594102.

[37] S. Khurshid, C. Reeder, L. X. Harrington, P. Singh, G. Sarma, S. F. Friedman, P. Di Achille, N. Diamant, J. W. Cunningham, A. C. Turner, E. S. Lau, J. S. Haimovich, M. A. Al-Alusi,

X. Wang, M. D. R. Klarqvist, J. M. Ashburner, C. Diedrich, M. Ghadessi, J. Mielke, H. M. Eilken, A. McElhinney, A. Derix, S. J. Atlas, P. T. Ellinor, A. A. Philippakis, C. D. Anderson, J. E. Ho, P. Batra, S. A. Lubitz, Cohort design and natural language processing to reduce bias in electronic health records research, npj Digital Medicine 5 (2022) 1–14. URL: https://doi.org/10.1038/s41746-022-00590-0.

[38] O. J. Bear Don't Walk, IV, H. Reyes Nieva, S. S.-J. Lee, N. Elhadad, A scoping review of ethics considerations in clinical natural language processing, JAMIA Open 5 (2022) 1–11. URL: https://doi.org/10.1093/jamiaopen/ooac039.

[39] M. A. Ricci Lara, R. Echeveste, E. Ferrante, Addressing fairness in artificial intelligence for medical imaging, Nature Communications 13 (2022) 4581. URL: https://doi.org/10.1038/s41467-022-32186-3.

[40] A. J. Mitchell, D. Malone, C. C. Doebbeling, Quality of medical care for people with and without comorbid mental illness and substance misuse: systematic review of comparative studies, The British Journal of Psychiatry 194 (2009) 491–499. URL: https://doi.org/10.1192/bjp.bp.107.045732.

[41] J. J. Deferio, S. Breitinger, D. Khullar, A. Sheth, J. Pathak, Social determinants of health in mental health care and research: a case for greater inclusion, Journal of the American Medical Informatics Association 26 (2019) 895–899. URL: https://doi.org/10.1093/jamia/ocz049.

[42] K. Mei, S. Fereidooni, A. Caliskan, Bias Against 93 Stigmatized Groups in Masked Language Models and Downstream Sentiment Classification Tasks, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 1699–1710. URL: https://doi.org/10.1145/3593013.3594109.

[43] B. Hutchinson, V. Prabhakaran, E. Denton, K. Webster, Y. Zhong, S. Denuyl, Social Biases in NLP Models as Barriers for Persons with Disabilities, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5491–5501. URL: https://doi.org/10.18653/v1/2020.acl-main.487.

[44] G. A. del Carmen, S. Stapleton, M. Qadan, M. G. del Carmen, D. Chang, Does the Day of the Week Predict a Cesarean Section? A Statewide Analysis, Journal of Surgical Research 245 (2020) 288–294. URL: https://10.1016/j.jss.2019.07.027.

[45] J. S. Mill, Utilitarianism, in: Seven Masterpieces of Philosophy, Routledge, London, UK, 2016, pp. 329–375.

[46] T. L. Beauchamp, J. F. Childress, Principles of biomedical ethics, 7th ed ed., Oxford University Press, New York, 2013.

[47] J. Lalor, Y. Yang, K. Smith, N. Forsgren, A. Abbasi, Benchmarking Intersectional Biases in NLP, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 3598–3609. URL: https://doi.org/10.18653/v1/2022.naacl-main.263.

[48] M. A. Wójcik, Assessing the Legality of Using the Category of Race and Ethnicity in Clinical Algorithms - the EU Anti-discrimination Law Perspective, in: Proceedings of the 2nd European Workshop on Algorithmic Fairness, volume 3442, CEUR Workshop

Proceedings, Winterthur, Switzerland, 2023, pp. 1–17.
[49] A. A. Abdu, I. V. Pasquetto, A. Z. Jacobs, An Empirical Analysis of Racial Categories in the Algorithmic Fairness Literature, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 1324–1333. URL: https://doi.org/10.1145/3593013.3594083.
[50] H. Devinney, J. Björklund, H. Björklund, Theories of "Gender" in NLP Bias Research, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 2083–2102. URL: https://doi.org/10.1145/3531146.3534627. doi:10.1145/3531146.3534627.
[51] J. R. Brown, I. M. Ricket, R. M. Reeves, R. U. Shah, C. A. Goodrich, G. Gobbel, M. E. Stabler, A. M. Perkins, F. Minter, K. C. Cox, C. Dorn, J. Denton, B. E. Bray, R. Gouripeddi, J. Higgins, W. W. Chapman, T. MacKenzie, M. E. Matheny, Information Extraction From Electronic Health Records to Predict Readmission Following Acute Myocardial Infarction: Does Natural Language Processing Using Clinical Notes Improve Prediction of Readmission?, Journal of the American Heart Association 11 (2022) e024198. URL: https://doi.org/10.1161/JAHA.121.024198, publisher: Wiley.
[52] B. D. Wissel, H. M. Greiner, T. A. Glauser, F. T. Mangano, D. Santel, J. P. Pestian, R. D. Szczesniak, J. W. Dexheimer, Investigation of bias in an epilepsy machine learning algorithm trained on physician notes, Epilepsia 60 (2019) e93–e98. URL: https://doi.org/10.1111/epi.16320.
[53] E. Wellesley Wesley, I. Patel, G. Kadra-Scalzo, M. Pritchard, H. Shetty, M. Broadbent, A. Segev, R. Patel, J. Downs, J. H. MacCabe, R. D. Hayes, D. F. de Freitas, Gender disparities in clozapine prescription in a cohort of treatment-resistant schizophrenia in the South London and Maudsley case register, Schizophrenia Research 232 (2021) 68–76. URL: https://doi.org/10.1016/j.schres.2021.05.006.
[54] M. Y. Yang, G. H. Kwak, T. Pollard, L. A. Celi, M. Ghassemi, Evaluating the Impact of Social Determinants on Health Prediction in the Intensive Care Unit, in: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, AIES '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 333–350. URL: https://doi.org/10.1145/3600211.3604719.
[55] D. Adjaye-Gbewonyo, R. A. Bednarczyk, R. L. Davis, S. B. Omer, Using the Bayesian Improved Surname Geocoding Method (BISG) to Create a Working Classification of Race and Ethnicity in a Diverse Managed Care Population: A Validation Study, Health Services Research 49 (2014) 268–283. URL: https://doi.org/10.1111/1475-6773.12089.
[56] L. Cheng, I. O. Gallegos, D. Ouyang, J. Goldin, D. Ho, How Redundant are Redundant Encodings? Blindness in the Wild and Racial Disparity when Race is Unobserved, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 667–686. URL: https://doi.org/10.1145/3593013.3594034, 0 citations (Crossref) [2023-12-07].
[57] R. Agarwal, M. Bjarnadottir, L. Rhue, M. Dugas, K. Crowley, J. Clark, G. Gao, Addressing algorithmic bias and the perpetuation of health inequities: An AI bias aware framework, Health Policy and Technology 12 (2023) 100702. URL: https://doi.org/10.1016/j.hlpt.2022.100702.
[58] M. Mccradden, O. Odusi, S. Joshi, I. Akrout, K. Ndlovu, B. Glocker, G. Maicas, X. Liu,

M. Mazwi, T. Garnett, L. Oakden-Rayner, M. Alfred, I. Sihlahla, O. Shafei, A. Goldenberg, What's fair is… fair? Presenting JustEFAB, an ethical framework for operationalizing medical ethics and social justice in the integration of clinical machine learning: JustEFAB, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 1505–1519. URL: https://dl.acm.org/doi/10.1145/3593013.3594096. doi:10.1145/3593013.3594096, 1 citations (Crossref) [2023-12-07].

[59] Y. Xiao, S. Lim, T. J. Pollard, M. Ghassemi, In the Name of Fairness: Assessing the Bias in Clinical Record De-identification, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 123–137. URL: https://doi.org/10.1145/3593013.3593982.

[60] W. Boag, H. Kané, S. Rawat, J. Wei, A. Goehler, A Pilot Study in Surveying Clinical Judgments to Evaluate Radiology Report Generation, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 458–465. URL: https://doi.org/10.1145/3442188.3445909.

[61] J. Guan, R. Li, S. Yu, X. Zhang, A Method for Generating Synthetic Electronic Medical Record Text, IEEE/ACM transactions on computational biology and bioinformatics 18 (2021) 173–182. URL: https://doi.org/10.1109/TCBB.2019.2948985.

[62] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R. G. Mark, MIMIC-III, a freely accessible critical care database, Scientific Data 3 (2016) 160035. URL: https://www.nature.com/articles/sdata201635. doi:10.1038/sdata.2016.35, number: 1 Publisher: Nature Publishing Group.

[63] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, R. Mark, MIMIC-IV 2.1, 2022. URL: https://doi.org/10.13026/rrgf-xw32.

[64] I. Neamatullah, M. M. Douglass, L.-w. H. Lehman, A. Reisner, M. Villarroel, W. J. Long, P. Szolovits, G. B. Moody, R. G. Mark, G. D. Clifford, Automated de-identification of free-text medical records, BMC Medical Informatics and Decision Making 8 (2008) 32. URL: https://doi.org/10.1186/1472-6947-8-32.

[65] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, S. Horng, MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports, Scientific Data 6 (2019) 317. URL: https://doi.org/10.1038/s41597-019-0322-0.

[66] E. Lehman, V. Lialin, K. E. Legaspi, A. J. Sy, P. T. Pile, N. R. Alberto, R. R. Ragasa, C. V. Puyat, M. K. Taliño, I. R. Alberto, P. G. Alfonso, D. Moukheiber, B. Wallace, A. Rumshisky, J. Liang, P. Raghavan, L. A. Celi, P. Szolovits, Learning to Ask Like a Physician, in: T. Naumann, S. Bethard, K. Roberts, A. Rumshisky (Eds.), Proceedings of the 4th Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Seattle, WA, 2022, pp. 74–86. URL: https://doi.org/10.18653/v1/2022.clinicalnlp-1.8.

[67] B. Boecking, N. Usuyama, S. Bannur, D. C. Castro, A. Schwaighofer, S. Hyland, M. Wetscherek, T. Naumann, A. Nori, J. Alvarez-Valle, H. Poon, O. Oktay, Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing, in: S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, T. Hassner (Eds.), Computer Vision – ECCV 2022, Lecture Notes in Computer Science, Springer Nature Switzerland, Cham, 2022, pp. 1–21.

URL: https://doi.org/10.1007/978-3-031-20059-5_1.
[68] Y. Gao, D. Dligach, T. Miller, S. Tesch, R. Laffin, M. M. Churpek, M. Afshar, Hierarchical Annotation for Building A Suite of Clinical Natural Language Processing Tasks: Progress Note Understanding, LREC ... International Conference on Language Resources & Evaluation : [proceedings]. International Conference on Language Resources & Evaluation 2022 (2022) 5484–5493. URL: https://aclanthology.org/2022.lrec-1.587.
[69] J. D. Osborne, T. O'Leary, A. Mudano, J. Booth, G. Rosas, G. Peramsetty, A. Knighton, J. Foster, K. Saag, M. I. Danila, Gout Emergency Department Chief Complaint Corpora, 2020. URL: https://doi.org/10.13026/96v3-dw72.
[70] Y. Huang, J. Guo, W. T. Donahoo, Z. Fan, Y. Lu, W.-H. Chen, H. Tang, L. Bilello, E. A. Shenkman, J. Bian, Developing A Fair Individualized Polysocial Risk Score (iPsRS) for Identifying Increased Social Risk of Hospitalizations in Patients with Type 2 Diabetes (T2D), 2023. URL: http://doi.org/10.48550/arXiv.2309.02467.
[71] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, H. E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals, Circulation [Online] 101 (2000) e215–e220.
[72] E. Farre Maduell, S. Lima-Lopez, S. A. Frid, A. Conesa, E. Asensio, A. Lopez-Rueda, H. Arino, E. Calvo, M. J. Bertran, M. A. Marcos, M. Nofre Maiz, L. Tañá Velasco, A. Marti, R. Farreres, X. Pastor, X. Borrat Frigola, M. Krallinger, CARMEN-I: A resource of anonymized electronic health records in Spanish and Catalan for training and testing NLP tools, 2023. URL: https://doi.org/10.13026/bxrx-y344.
[73] H. Dias, A. H. D. P. d. Ulbrich, BRATECA (Brazilian Tertiary Care Dataset): a Clinical Information Dataset for the Portuguese Language, 2022. URL: https://doi.org/10.13026/ay8n-qf21.