ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Factors associated with social determinants of health mentions in PubMed clinical case reports from 1975 to 2022: A natural language processing analysis

Julio Bonis*, Veysel Kocaman, and David Talby

John Snow Labs Inc., Delaware, United States of America

Abstract

Social determinants of health (SDoH) significantly influence health outcomes, accounting for nearly 40% of such outcomes globally. These determinants, pivotal in understanding health disparities, are insufficiently documented in clinical settings and academic clinical narratives. To address this gap, we examined clinical case reports from PubMed (1975–2022) to identify mentions of six specific SDoH, employing a pre-trained named-entity recognition (NER) model from Spark natural language processing (NLP). Multivariate logistic regression was utilized to investigate associations between article characteristics and the documentation of SDoH. From 463,546 reports, 4.4% mentioned SDoH, with race/ethnicity being the most dominant mention. Race/ethnicity was often cited by sub-Saharan African authors (adjusted odds ratio [AOR]: 4.47) and in general medicine (AOR: 2.18). Marital status mentions appeared predominantly in psychiatry (AOR: 2.60) and gynecology (AOR: 2.47). Sexual orientation mentions were correlated with infectious diseases (AOR: 25.00) and varied by authorship regions, with stronger associations observed in South America (AOR: 4.04) and North America (AOR: 2.15), and comparatively weaker associations noted in the Indian subcontinent and the Middle East (AOR: 0.16). Immigrant status mentions were closely related to infectious diseases (AOR: 4.51), gynecology (AOR: 4.25), and certain geographies. Homelessness mentions were more prominent in forensic medicine (AOR: 14.92) and in both infections (AOR: 6.36) and mental disorders (AOR: 5.80). Spiritual belief mentions were more prominent with sub-Saharan authors (AOR: 9.17) and psychiatry (AOR: 7.61). SDoH mentions in medical literature were also determined by the diagnosis, cultural background, and journal type. The limited SDoH registration emphasized their overlooked significance. Disproportionate emphasis on specific relationships, such as sexual orientation with infectious diseases, can perpetuate biases and stereotypes. Innovative tools such as Spark NLP offer promise in advancing research using electronic health records (EHRs), but a standardized approach to SDoH reporting and vigilant AI training is crucial for unbiased health-care analysis.

Keywords: Social determinants of health; Natural language processing; Clinical case reports; Ethnicity; Marital status; Immigrant status; Homeless; Spiritual beliefs

***Corresponding author:**
Julio Bonis
(julio@johnsnowlabs.com)

Citation: Bonis J, Kocaman V, Talby D. Factors associated with social determinants of health mentions in PubMed clinical case reports from 1975 to 2022: A natural language processing analysis. *Artif Intell Health.* 2024;1(2): 117-131. doi: 10.36922/aih.2737

Received: January 14, 2024

Accepted: March 18, 2024

Published Online: April 17, 2024

1. Introduction

Social determinants of health (SDoH) are fundamental conditions that underpin the health disparities experienced by individuals globally. These determinants are the circumstances in which people are born, grow, work, and live, and they encompass factors such as socioeconomic status, housing, food security, and exposure to violence or stress.[1,2] Notably, these conditions have been proven to shape health outcomes to such an extent that up to 40% of health outcomes are attributed to SDoH challenges.[3,4]

Significantly, SDoH not only impacts health outcomes but also has discernible effects on health-care utilization. For instance, unmet social needs, a facet of SDoH, have been tied to clinical outcomes such as uncontrolled diabetes,[5] hypertension,[6] and increased hospital readmissions for heart failure.[7] There is also evidence suggesting that moving from a high-poverty neighborhood to one with lower poverty levels can lead to reductions in conditions such as extreme obesity and diabetes, emphasizing the role of environmental factors on health.[8]

Given the undeniable influence of SDoH on health, there have been initiatives to incorporate SDoH screening into health-care delivery, with proposals to standardize the methods for capturing this information in electronic health records (EHRs).[9] Advocates believe that documenting SDoH systematically at the point of care would bolster the identification of patients' risk factors and streamline referrals to social services, fostering a more holistic approach to patient care.[10,11]

However, the current reality paints a different picture. Despite the evident significance of SDoH, they remain underrepresented in clinical documentation. Recent studies have indicated that a mere 2% of patients visiting community health centers had at least one documented SDoH,[12] a figure that was confirmed by the analysis of the ICD10 codes in different studies.[13,14] Moreover, another study examining over a million unique patient EHRs found that only a small percentage contained mentions of social isolation, housing issues, or financial strain,[15] a finding that has been replicated in other studies.[16] However, other analyses conducted in the primary care context have reported slightly higher proportions i.e., 7% of patients with SDoH documented in Spain[17] and 4% to 18% in the United States (US).[18] These findings indicate that utilizing EHRs for SDoH documentation is insufficient, and a systemic approach involving education, policy redesign, and incentives might be necessary to boost documentation.[9]

These findings are concerning as a discrepancy in SDoH documentation could be indicative of a broader oversight in clinical decision-making. Within the domain of medical literature, clinical case reports serve as a reflection of the priorities and perspectives of health-care professionals. The choices they make in detailing specific patient information — what they choose to include or exclude—offer insights into what they deem significant or irrelevant. As such, the inclusion or omission of SDoH in these published reports can act as a barometer of their importance within the health-care community. By analyzing the frequency and context of SDoH mentions in these clinical cases, one can gauge the weight and significance attributed to these factors by health-care professionals when communicating notable clinical findings to a wider scientific audience.

Natural language processing (NLP) has become an indispensable tool in the medical domain, revolutionizing the extraction and analysis of complex data from clinical texts and patient records. Recent publications[19,20] highlight the crucial role of NLP in identifying, categorizing, and analyzing health-related information from unstructured content as clinical narratives. The advancements in NLP technologies, such as context-aware models like Bidirectional Encoder Representations from Transformers (BERT)[21] and BioBERT,[22] have dramatically enhanced our ability to process vast datasets, thereby transforming traditional health-care data analysis methods.[23-26] These innovations offer deeper insights into the prevalence and impact of SDoH, previously obscured in clinical documentation.[27] For instance, research has demonstrated that NLP-based systems can identify clinical events with significantly higher precision and sensitivity compared to traditional methods. One study demonstrated that an NLP system identified approximately four times as many clinical events as standard approaches, with a positive predictive value (PPV) of 74%, a stark improvement over the 31% PPV of methods relying solely on diagnostic codes.[28] In another study, the precision of selected cases increased from 46% to 86% after incorporating NLP methods that followed structured-based case selection with a sensitivity of 77%.[29] These examples highlight the transformative impact of NLP in enhancing the detection and characterization of SDoH and clinical events from medical narratives, enabling a more nuanced and comprehensive analysis of health-care data.

Our study utilizes advanced NLP technology to meet the need for improved documentation and understanding of SDoH in clinical settings. We investigated factors influencing the mention of SDoH in publicly available clinical case reports and how this knowledge could inform the development of more effective policies for SDoH reporting. In addition, our analysis identified potential stereotypes or discrimination in artificial intelligence (AI)

models trained in the medical literature. We believe that our research adds to the discussion on SDoH, which could consequently enhance AI tools and policies for unbiased reporting of these determinants.

2. Methods

We obtained the latest annual PubMed baseline (available on September 1, 2023) through File Transfer Protocol (FTP) and parsed the search results to exclusively display publications tagged as "Clinical Case Report," yielding a total of 1,643,513 reports. We refined the search for articles published from January 1, 1975, to December 31, 2022. In addition, we employed a set of regular expressions to only include papers with abstracts that present a genuine clinical narrative about individual patients, rather than reports of aggregated case series. These were designed to pinpoint abstracts that mention both the age and gender of a single patient, resulting in the identification of 463,546 relevant articles (Figure 1).

To delineate the content of each article, we utilized a deep learning-based sentence boundary detection model[25,30] and produced a list of sentences for every article. Our focus was strictly on sentences that mentioned the patients' age and gender and identified using the same set of regular expressions. These sentences were then input into a pre-trained named-entity recognition (NER) model from John Snow Labs (JSL), designed to identify mentions associated with various SDoH and based on a proprietary fine-tuned BERT architecture.[31,32]

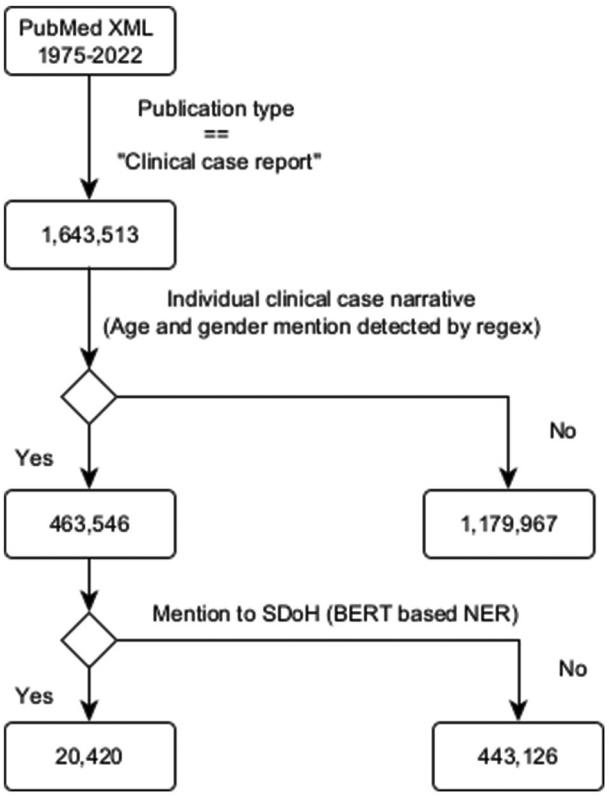

Figure 1. Workflow diagram illustrating the selection process of clinical case reports. The figure was created with yEd.
Abbreviations: BERT: Bidirectional Encoder Representations from Transformers for Biomedical Text Mining; NER: Named-entity recognition; SDoH: Social determinants of health; XML: Extensible markup language.

The accuracy of the model was assessed with an external dataset from JSL, encompassing 9,743 sentences and 198,698 tokens with manually annotated mentions to SDoH, namely race/ethnicity (n = 72), sexual orientation (n = 20), marital status (n = 193), housing (n = 371), population subgroup (n = 19), and spiritual beliefs (n = 90). This external test also compared the outcomes to generative pre-trained transformer (GPT)-3.5[33] and GPT-4.[34] In addition, an internal validation reviewed the precision for each SDoH entity found by the model in the PubMed dataset used in this study.

Besides the formal evaluation that considered the specific assertions of entities, our internal analysis prioritized identifying factors linked to SDoH mentions in clinical narratives. Hence, it was unnecessary to delve into the precise details or assertions regarding SDoH, such as a patient's marital status, whether they were married, unmarried, or if their marital status was unspecified. Our main interest was determining whether any SDoH mention, like marital status, was made, irrespective of its actual status or value. This method streamlined the extraction process by removing the need to navigate the intricacies associated with each SDoH status.

Consequently, our approach aligned with the study's objective to simply ascertain the occurrence of SDoH mentions within clinical documentation. Age and gender, used as selection criteria, were omitted from the SDoH evaluation. We targeted six specific SDoH, i.e., race/ethnicity, marital status, population group/immigrant status, sexual orientation, spiritual beliefs, and housing/homelessness, and analyzed them based on recall, precision, exclusion of individual behavior determinants not essentially social, and minimum corpus occurrence of 50 matches.

The journals' geographic origins were identified from PubMed records, and the first author's geographic origin was obtained from their reported affiliation. The main diagnosis was obtained from PubMed's Medical Subject Headings (MeSH) codes corresponding to disease or mental condition categories. Only root primary disease categories (e.g., respiratory tract, neurological, and mental conditions) were used during the analysis.

To analyze the link between article features and SDoH mentions, we conducted six logistic regression analyses using the Python package statsmodels 0.14.0 to gauge the adjusted odds ratio (AOR) for each article trait. We also employed a stepwise additive method,[35] where features that could enhance the likelihood of the model were sequentially incorporated with a *P*-value threshold of 0.001 for the likelihood ratio test.

3. Results

3.1. Study population and data inclusion

We analyzed a comprehensive dataset comprising 463,546 clinical case reports indexed in Medline from 1975 through 2022. The distribution of the articles based on four key information (author's geographic region, journal's geographic region, journal specialty, and clinical diagnosis) is displayed in Table 1.

3.2. Recall and precision of identifying mentions of the social determinants of health

In our corpus analysis, the SDoH identification precisions were 99.3% (95% confidence interval [CI]: 99.2 – 99.4%) for race/ethnicity, 90.2% (95% CI: 88.8 – 91.4%) for marital status, 90.8% (95% CI: 86.9–93.6%) for population group, 97.4% (95% CI: 95.6 – 98.4%) for sexual orientation, 100% (95% CI: 94.6 – 100%) for housing, and 98.4% (95% CI: 91.7 – 99.7%) for spiritual beliefs.

During external validation, the precision results were 97.4% (95% CI: 86.5 – 99.5%) for race/ethnicity, 100% (95% CI: 92.3 – 100%) for marital status, 88.9% (95% CI: 56.5 – 98.0%) for population group, 93.8% (95% CI: 71.7 – 98.9%) for sexual orientation, 98.6% (95% CI: 92.3 – 99.7%) for housing, and 83.0% (95% CI: 70.8 – 90.8%) for spiritual beliefs.

The recalls in the external validation were 90.2% (95% CI: 77.5 – 96.1%) for race/ethnicity, 97.9% (95% CI: 88.9 – 99.6%) for marital status, 88.9% (95% CI: 56.5 – 98.0%) for population group, 100% (95% CI: 79.6 – 100%) for sexual orientation, 85.2% (95% CI: 75.9 – 91.37%) for housing, and 83.0% (95% CI: 70.8 – 90.8%) for spiritual beliefs.

In our analysis comparing the recall and precision of the JSL SDoH-NER model with those of zero-shot learning (i.e., GPT-3.5 and GPT-4), both JSL and GPT-4 displayed comparable results. Notably, some differences were evident: JSL outperformed GPT-4 in precision for marital status (p = 0.005; GPT-4 scored 82.9%; 95% CI: 67.3–91.9%) and housing (p < 0.001; GPT-4 scored 82.9%; 95% CI: 67.3–91.9%). The results of this comparison are detailed in Figures S1 and S2.

3.3. Prevalence of social determinants of health mentions

Among the total case reports examined, 20,420 (4.4%) case reports included references to at least one SDoH category. A breakdown revealed that 17,765 case reports specifically mentioned race/ethnicity, followed by 1,991 articles that discussed marital status, 524 on sexual orientation, 284 on immigrant status, 63 on spiritual beliefs, and 60 on homelessness. The mean and confidence intervals of the mentioned rates within the study period are summarized in Table 2.

The analysis of the proportion of clinical cases reporting SDoH within the study period indicated a statistically significant association between publication year and race/ethnicity (P < 0.001), sexual orientation (P < 0.001), and homelessness (P < 0.001), respectively. Notably, there was a peak of sexual orientation mentions from 1980 to 1995, and we hypothesized that this could be related to the rise of acquired immunodeficiency syndrome (AIDS) cases, as depicted in Figure S3. There was also a prominent increase in race/ethnicity mentions between 2011 and 2013 (Figure S4) and a less evident but statistically significant increase in homelessness mentions since 1990.

3.4. Factors associated with reporting social determinants of health

3.4.1. Race/ethnicity

Significant associations were observed between the author's geographic origins and the frequency of race/ethnicity mentions. Authors from sub-Saharan Africa were most likely to discuss race/ethnicity (AOR: 4.47; 95% CI: 3.96 – 5.04), followed by the Caribbean (AOR: 3.31; 95% CI: 2.24 – 4.89), Southeast Asia (AOR: 2.89; 95% CI: 2.58 – 3.25), East Asia (AOR: 2.00; 95% CI: 1.90 – 2.09), and North America (AOR: 1.77; 95% CI: 1.68 – 1.86). Conversely, authors from the Indian subcontinent (AOR: 0.69; 95% CI: 0.62 – 0.76) and Middle East (AOR: 0.77; 95% CI: 0.70 – 0.84) were less inclined to mention race/ethnicity in their case reports.

The journal's geographic region also exerted an independent influence on race/ethnicity mentions. Journals originating from Australia-Oceania (AOR: 1.34; 95% CI: 1.17 – 1.53) and Western Europe (AOR: 1.30; 95% CI: 1.18 – 1.43) were slightly more prone to include race/ethnicity. In contrast, journals from East Asia (AOR: 0.48; 95% CI: 0.43 – 0.54), Eastern Europe (AOR: 0.54; 95% CI: 0.45 – 0.64), and South America (AOR: 0.55; 95% CI: 0.43 – 0.69) had much fewer race/ethnicity mentions than expected.

Table 1. Information on the analyzed articles (*n*=463546)

Information	Distribution	Number of articles	Percentage distribution of articles (%)
Author's geographic region	Known	334666	72.20
	East Asia	95527	28.54
	Western Europe	94950	28.37
	North America	72892	21.78
	Middle East	23631	7.06
	Indian subcontinent	13809	4.13
	Eastern Europe	9079	2.71
	South America	8299	2.48
	Australia and Oceania	6283	1.88
	Southeast Asia	3383	1.01
	Sub-Saharan Africa	2688	0.80
	North Africa	2440	0.73
	Central America	1395	0.42
	Caribbean	265	0.08
	Central Asia	25	0.01
	Unknown	128880	27.80
Journal's geographic region	Known	462600	99.80
	Western Europe	196878	42.56
	North America	150489	32.53
	East Asia	72101	15.59
	Eastern Europe	11157	2.41
	Australia and Oceania	8674	1.88
	Indian subcontinent	6780	1.47
	Middle East	6470	1.40
	South America	3759	0.81
	Sub-Saharan Africa	3101	0.67
	Southeast Asia	1657	0.36
	North Africa	617	0.13
	Central America	612	0.13
	Caribbean	305	0.07
	Unknown	946	0.20
Journal specialty	Known	423452	91.35
	General medicine	85521	20.20
	Surgery	77849	18.38
	Neurology	30533	7.21
	Oncology	23319	5.51
	Pediatrics	19518	4.61
	Cardiology	19393	4.58
	Dermatology	17516	4.14
	Pathology	17205	4.06
	Ophthalmology	15254	3.60
	Gastroenterology	12554	2.96
	Laboratory	12123	2.86

(*Cont'd...*)

Table 1. (*Continued*)

Information	Distribution	Number of articles	Percentage distribution of articles (%)
	Radiology	11792	2.78
	Urology	11641	2.75
	Gynecology	9354	2.21
	Infectiology	9190	2.17
	Traumatology	8028	1.90
	Hematology	6968	1.65
	Anesthesiology	6398	1.51
	Endocrinology	4996	1.18
	Neurology	4911	1.16
	Rheumatology	3687	0.87
	Nephrology	3555	0.84
	Psychiatry	3273	0.77
	Dentistry	2655	0.63
	Forensic	2332	0.55
	Public Health	1165	0.28
	Rehabilitation	1140	0.27
	Genetics	931	0.22
	Allergy	651	0.15
	Unknown	40094	8.65
Diagnosis	Neoplasms	154185	33.26
	Pathological signs and symptoms	117438	25.33
	Nervous system diseases	83899	18.10
	Infections	68717	14.82
	Cardiovascular diseases	67711	14.61
	Digestive system diseases	40355	8.71
	Musculoskeletal diseases	38527	8.31
	Urogenital diseases	37470	8.08
	Respiratory tract diseases	31740	6.85
	Hemic and lymphatic diseases	30350	6.55
	Skin and connective tissue diseases	22786	4.92
	Nutritional and metabolic diseases	20015	4.32
	Wounds and injuries	19674	4.24
	Eye diseases	19475	4.20
	Congenital, hereditary, and neonatal diseases	15903	3.43
	Stomatognathic diseases	9776	2.11
	Endocrine system diseases	9768	2.11
	Mental disorders	9109	1.97
	Chemically-induced disorders	7722	1.67
	Immune system diseases	7054	1.52
	Otorhinolaryngologic diseases	4339	0.94
	Occupational diseases	914	0.20
	Animal diseases	394	0.08
	Disorders of environmental origin	2	0.00

Note: Percentages of known characteristics are expressed relative to the total number of known articles; the cumulative percentage of diagnoses is more than 100% as a single article can have one or more assigned diagnoses; the list of diagnoses is based on the Medical Subject Headings (MeSH).

Table 2. Average SDoH mentions from clinical case reports (*n*=463546) between 1975 and 2022

SDoH	SDoH mentions (95% CI)
Race/ethnicity	383.24 (377.71–388.77)
Marital status	42.95 (41.06–44.83)
Sexual orientation	11.30 (10.34–12.27)
Immigrant status	6.13 (5.41–6.84)
Spiritual beliefs	1.36 (1.02–1.69)
Homelessness	1.29 (0.97–1.62)

Abbreviations: CI: Confidence interval; SDoH: Social determinants of health.

The specialty of the journal significantly influenced the likelihood of race/ethnicity mentions. Case reports in general medicine were the most likely to include race/ethnicity (AOR: 2.18; 95% CI: 2.08 – 2.29), followed by laboratory medicine (AOR: 2.10; 95% CI: 1.94 – 2.28), dentistry (AOR: 1.82; 95% CI: 1.55 – 2.13), and psychiatry (AOR: 1.82; 95% CI: 1.56 – 2.13). A moderate tendency to mention race/ethnicity was also observed in other journal specialties (AOR: 1.37 – 1.97) (Table S1). Surgical specialties were generally less likely to mention race/ethnicity. These included anesthesiology (AOR: 0.27; 95% CI: 0.20 – 0.37), urology (AOR: 0.48; 95% CI: 0.40 – 0.56), traumatology (AOR: 0.59; 95% CI: 0.50 – 0.70), and general surgery (AOR: 0.61; 95% CI: 0.57 – 0.65). Rehabilitation (AOR: 0.31; 95% CI: 0.18 – 0.54) and radiology (AOR: 0.40; 95% CI: 0.35 – 0.47) displayed a strong tendency against reporting race/ethnicity in their clinical cases. Some journal specialties, namely cardiology (AOR: 0.63; 95% CI: 0.56 – 0.72), pneumology (AOR: 0.75; 95% CI: 0.61 – 0.92), and neurology (AOR: 0.79; 95% CI: 0.72 – 0.87), were slightly less inclined to include this information in their clinical case reports.

Finally, the primary diagnosis of the clinical case was also correlated with the likelihood of race/ethnicity mentions, although less strongly than the other variables. Hematological, eye, stomatognathic, metabolic, skin diseases, and infections were significantly associated with slightly higher mentions of race/ethnicity (AOR: 1.20 – 1.32). Conversely, occupational diseases, wounds and injuries, cardiovascular diseases, nervous system diseases, respiratory diseases, and digestive diseases were associated with fewer race/ethnicity mentions (AOR: 0.64 – 0.91).

Detailed information about the AOR of each factor associated with race/ethnicity mentions can be found in Figure 2 and Table S1.

3.4.2. Marital status

Mentions of marital status were notably correlated with several journal specialties such as psychiatry (AOR: 2.63; 95% CI: 1.97 – 3.51), gynecology (AOR: 2.45; 95% CI: 2.01 – 2.99), rehabilitation (AOR: 2.39; 95% CI: 1.31 – 4.35), and forensic medicine (AOR: 2.04; 95% CI: 1.32 – 3.17). Conversely, nephrology (AOR: 0.21; 95% CI: 0.09 – 0.51) and traumatology (AOR: 0.46; 95% CI: 0.26 – 0.79) displayed a pronounced negative correlation with mentions of marital status. Clinical cases pertaining to mental disorders (AOR: 2.14; 95% CI: 1.72 – 2.66) and urogenital diseases (AOR: 1.68; 95% CI: 1.47 – 1.91) were robustly associated with mentions of marital status. Authors from sub-Saharan Africa also exhibited a marked inclination to mention marital status (AOR: 1.98; 95% CI: 1.32 – 2.96).

Several other factors had associations with the likelihood of mentioning marital status, although more moderately. Clinical cases covering a broad spectrum of conditions, such as wounds, neoplasms, infections, digestive, hematological, skin, respiratory, metabolic, musculoskeletal, and nervous diseases, as well as those related to unspecific signs and symptoms, were linked with slightly fewer mentions of marital status (AOR: 0.51 – 0.77). Journals focusing on gastroenterology and general surgery (AOR: 0.53 – 0.74) also demonstrated a subtle association with reduced mentions of marital status.

Lastly, case reports published in the Indian subcontinent or authored by individuals from the Middle East, the Indian subcontinent, North Africa, and Southeast Asia were more inclined to mention marital status (AOR: 1.31 – 1.75). Further details on marital status mentions can be found in Figure 3 and Table S2.

3.4.3. Sexual orientation

The mention of sexual orientation was profoundly correlated with the diagnosis of infectious diseases (AOR: 25.00; 95% CI: 19.68 – 31.75). Other robustly associated factors include case reports published in South America (AOR: 4.04; 95% CI: 1.92 – 8.50) and North America (AOR: 2.15; 95% CI: 1.31 – 3.55). In contrast, journal specialties, such as pediatrics (AOR: 0.16; 95% CI: 0.07 – 0.39) and surgery (AOR: 0.46; 95% CI: 0.30 – 0.69), demonstrated a strong negative correlation with mentions of sexual orientation. A similar trend was also observed across a variety of diagnoses, including cardiovascular, musculoskeletal, and respiratory (AOR: 0.26 – 0.37).

Authors from the Indian subcontinent (AOR: 0.16; 95% CI: 0.04 – 0.63) and the Middle East (AOR: 0.16; 95% CI: 0.05 – 0.51) were considerably less inclined to mention sexual orientation. Conversely, authors from North America (AOR: 1.47; 95% CI: 1.13 – 1.91) and Western Europe (AOR: 1.46; 95% CI: 1.15 – 1.87) were more inclined to mention sexual orientation more frequently than the authors from other regions. Further details on

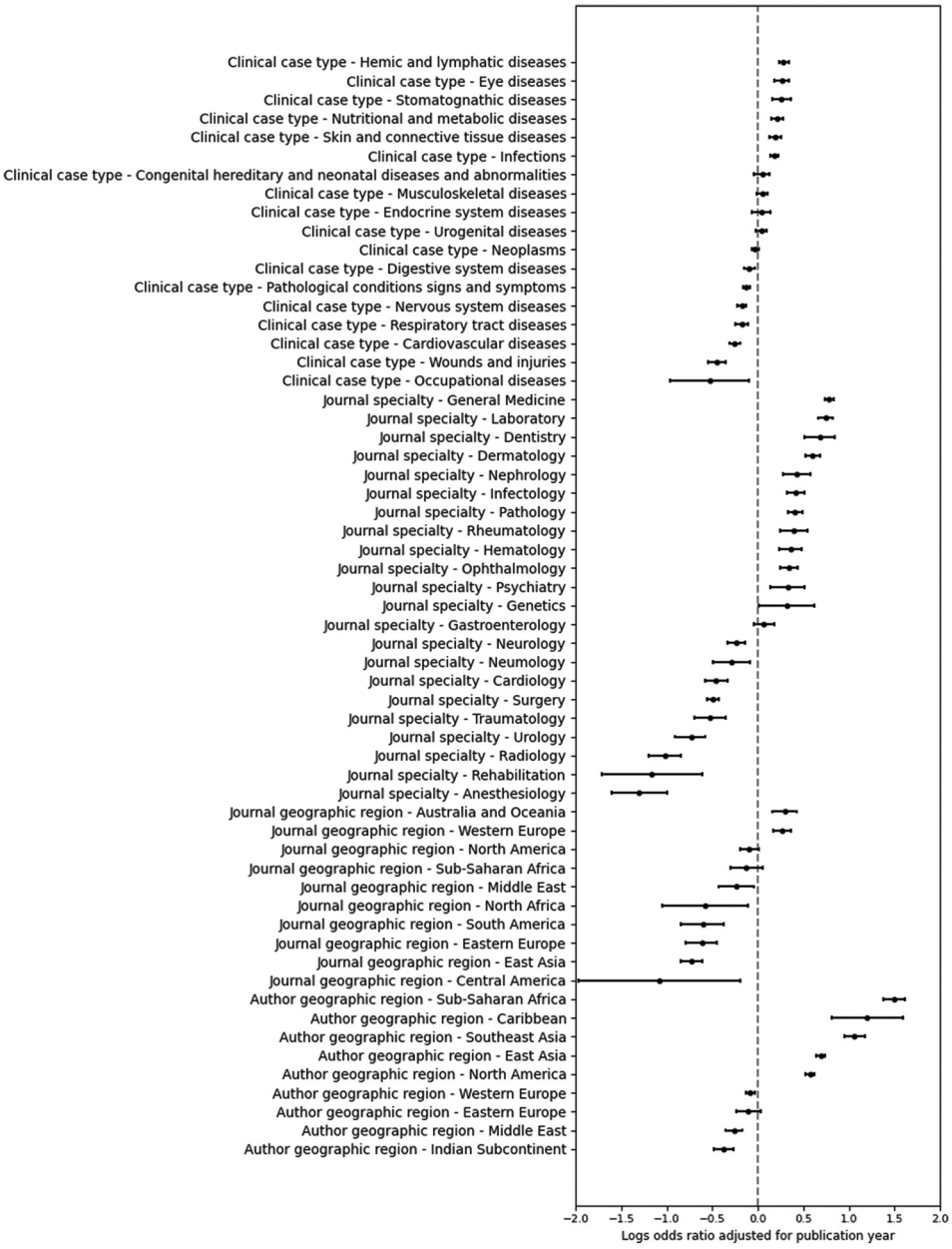

Figure 2. Adjusted odds ratios for the probability of mentioning race/ethnicity based on clinical case type, journal specialty, journal's geographic region, and author's geographic region. The figure was plotted with Matplotlib.

sexual orientation mentions can be found in Figure 4 and Table S3.

3.4.4. Immigrant status

Mentions of immigrant status were strongly associated with infectious diseases (AOR: 4.51; 95% CI: 3.53 – 5.77) and to a lesser extent, with mental disorders (AOR: 2.05; 95% CI: 1.14 – 3.71). Mentions of immigrant status were also positively and significantly associated with journals specializing in gynecology (AOR: 4.25; 95% CI: 2.64 – 6.82) and psychiatry (AOR: 3.94; 95% CI: 1.95 – 7.95), case reports published in the Middle East (AOR: 2.20; 95% CI: 1.19 – 4.07), and authors from Australia and Oceania (AOR: 2.17; 95% CI: 1.14 – 4.12).

Conversely, reduced mentions of immigrant status were associated with authors from the Indian subcontinent (AOR: 0.09; 95% CI: 0.01 – 0.62) and East Asia (AOR: 0.23; 95% CI: 0.12 – 0.45), case reports published in East Asia

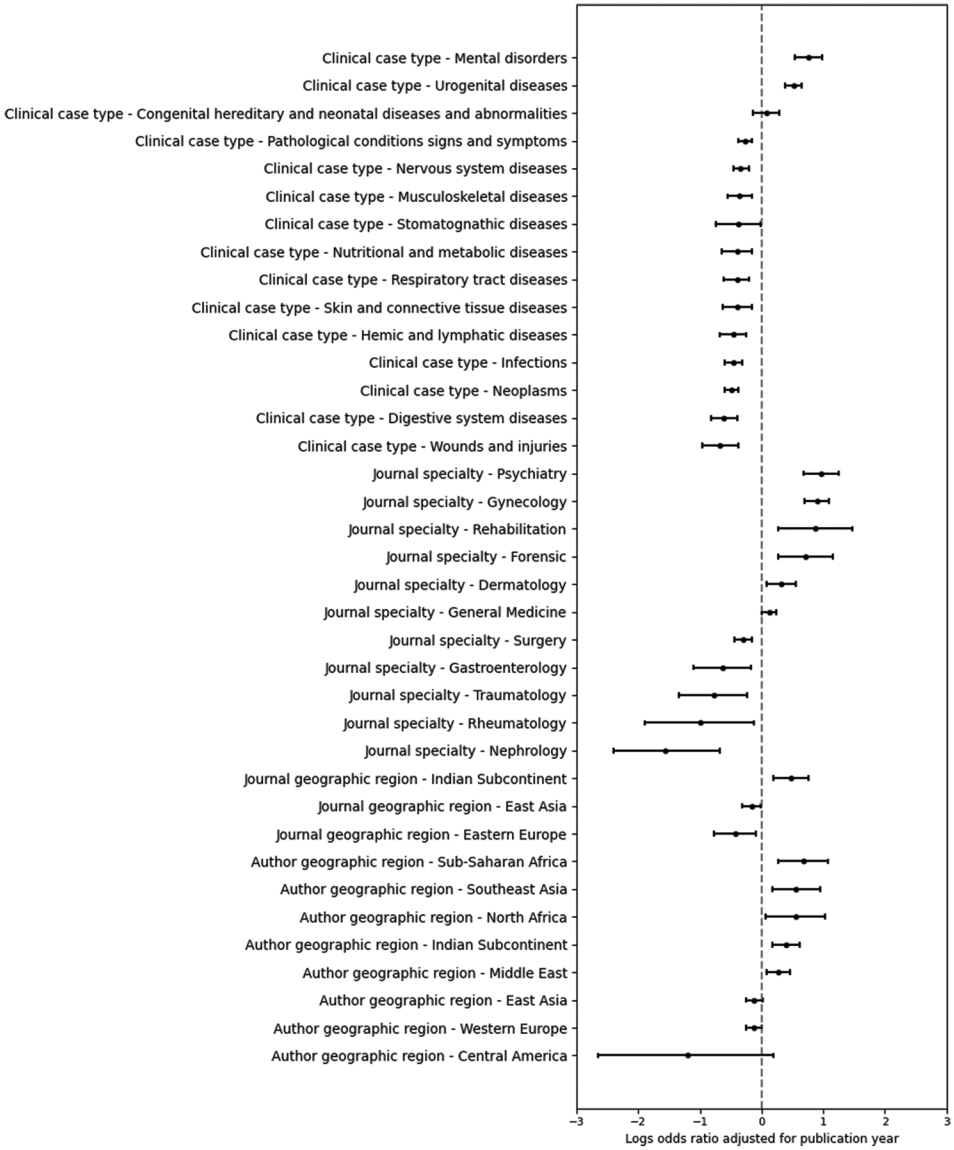

Figure 3. Adjusted odds ratios for the probability of mentioning marital status based on clinical case type, journal specialty, journal's geographic region, and author's geographic region. The figure was plotted with Matplotlib.

(AOR: 0.23; 95% CI: 0.15 – 0.65), and journals specializing in ophthalmology (AOR: 0.12; 95% CI: 0.02 – 0.92) and dermatology (AOR: 0.28; 95% CI: 0.09 – 0.90). Diagnoses pertaining to cardiovascular diseases (AOR: 0.43; 95% CI: 0.27 – 0.69) and neoplasms (AOR: 0.43; 95% CI: 0.31 – 0.63) also displayed marked negative associations with immigrant status mentions. Both general medicine journals (AOR: 1.75; 95% CI: 1.34 – 2.29) and authors from North America (AOR: 1.53; 95% CI: 1.17 – 2.01) demonstrated moderate positive associations with mentions of immigrant status. Further details on immigrant status mentions are available in Figure 5 and Table S4.

3.4.5. Homelessness

Mentions of homelessness were strongly associated with journals in the field of forensic medicine (AOR: 14.92; 95% CI: 5.48 – 40.64). Other strongly correlated factors included journals in the areas of pathology (AOR: 3.95; 95% CI: 1.39 – 11.28) and infectious diseases (AOR: 3.75; 95% CI: 1.77 – 7.94), publications from Eastern Europe (AOR: 4.76; 95% CI: 1.88 – 12.03), and diagnoses related to infections (AOR: 6.36; 95% CI: 3.57 – 11.32), mental disorders (AOR: 5.80; 95% CI: 2.26 – 14.89), and injuries (AOR: 4.73; 95% CI: 2.29 – 9.77). Further information on homelessness mentions is available in Figure 6 and Table S5.

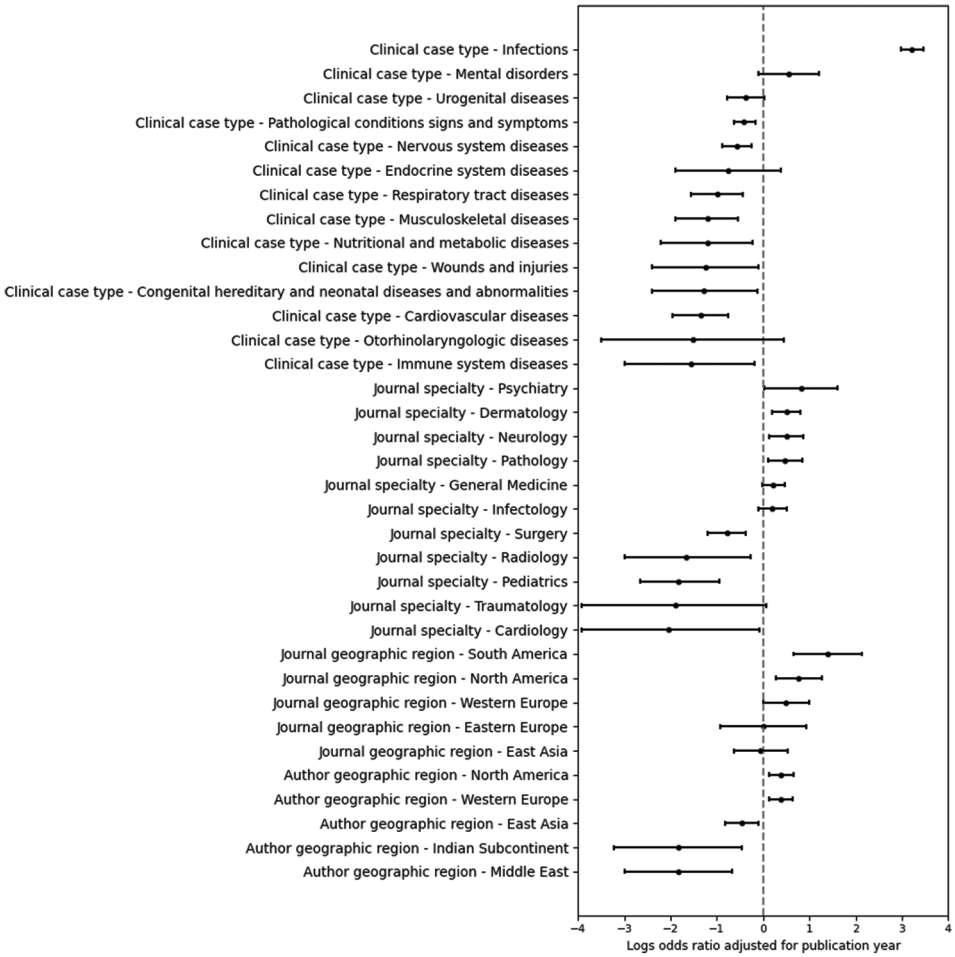

Figure 4. Adjusted odds ratios for the probability of mentioning sexual orientation based on clinical case type, journal specialty, journal's geographic region, and author's geographic region. The figure was plotted with Matplotlib.

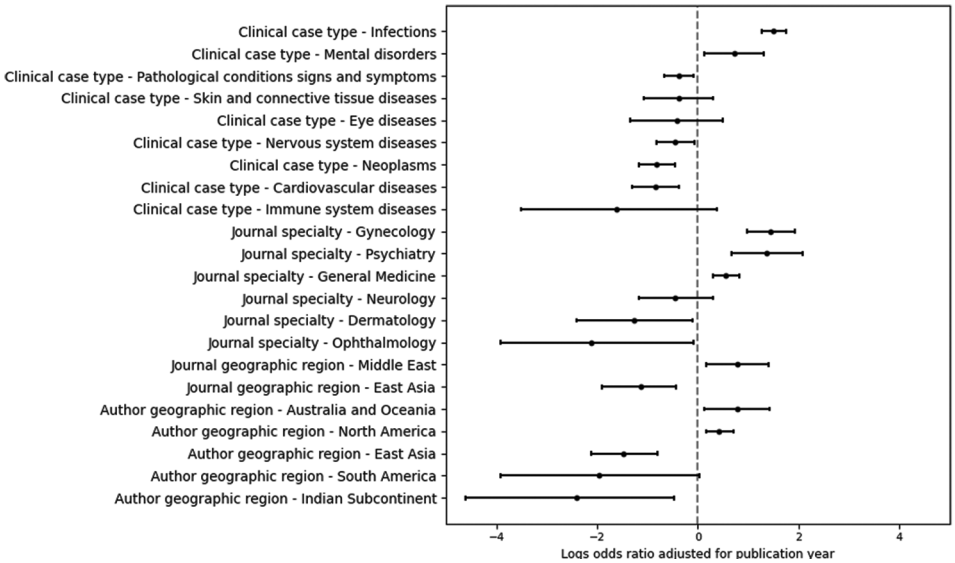

Figure 5. Adjusted odds ratios for the probability of mentioning immigrant status/population group based on clinical case type, journal specialty, journal's geographic region, and author's geographic region. The figure was plotted with Matplotlib.

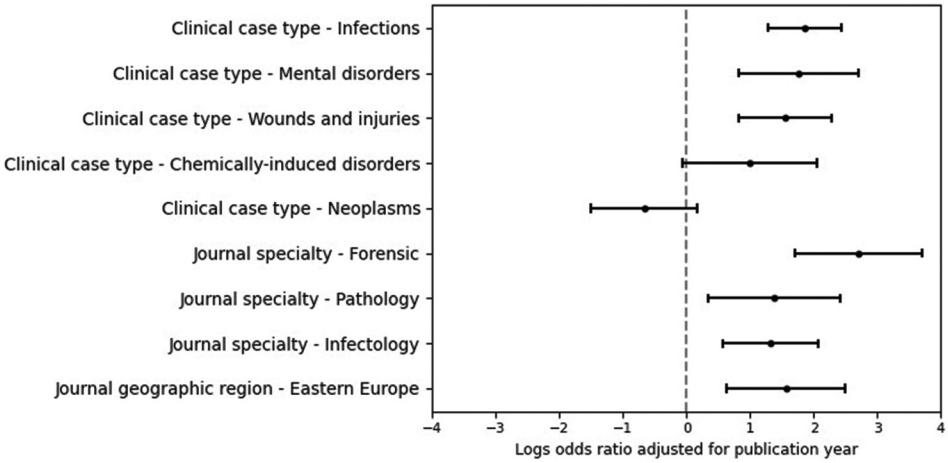

Figure 6. Adjusted odds ratios for the probability of mentioning homelessness/housing based on clinical case type, journal specialty, and journal's geographic region. The figure was plotted with Matplotlib.

3.4.6. Spiritual beliefs

Mentions of spiritual beliefs were strongly correlated with authors from sub-Saharan Africa (AOR: 9.17; 95% CI: 2.84 – 29.64) and the Indian subcontinent (AOR: 4.09; 95% CI: 1.83 – 9.15), journals in the field of psychiatry (AOR: 7.61; 95% CI: 2.93–19.79), publications from the Middle East (AOR: 5.05; 95% CI: 1.99 – 12.85), and clinical cases related to endocrine system diseases (AOR: 3.47; 95% CI: 1.38 – 8.68) and mental disorders (AOR: 3.05; 95% CI: 1.27 – 7.31). In contrast, journals in the field of surgery (AOR: 0.23; 95% CI: 0.06 – 0.96) and clinical cases related to neoplasms (AOR: 0.20; 95% CI: 0.08 – 0.50) were associated with lower probabilities of mentioning patients' spiritual beliefs. Further information on spiritual belief is included in Figure 7 and Table S6.

4. Discussion

4.1. Low prevalence of social determinants of health mentions

Our analysis revealed an uneven distribution of SDoH factors, such that three SDoH factors did not display a clear time-dependent trend. Regarding sexual orientation (Figure S3), a brief increase in mentions occurred in the 1980s, peaking at 40/10,000 case reports. However, the mentions of sexual orientation sharply decreased in the 2000s, leveling at 5/10,000 case reports. We theorized that this surge was associated with the AIDS/human immunodeficiency virus (HIV) outbreak in that period.

There was little variation in race/ethnicity mentions with time (until 2011), depicting steadiness at approximately 300/10,000 case reports (Figure S4). However, between 2011 and 2013, race/ethnicity mentions surged to nearly 550/10,000 case reports. This rate has persisted until 2022, indicating a lasting change in awareness or reporting about race/ethnicity. Nonetheless, further studies are warranted to investigate the reason for the observed trend.

Homelessness mentions displayed a slight increase, but the rate was only 1.29/10,000 case reports, contrasting with the estimated US 1-year homelessness prevalence – about 100 times higher.[36]

Collectively, the data revealed no consistent longitudinal SDoH reporting trends. Observable shifts were sporadic, brief, or tied to specific periods, highlighting the variability of SDoH in the medical literature.

4.2. Risk of biases in the social determinants of health

Our findings reported that diagnosis significantly affects SDoH mentions. Both individual cultural norms (reflected by the author's origins) and institutional policies (indicated by the journal's origins and specialties) impacted SDoH mention frequency. Notably, individual regional contexts exhibited distinct patterns when contrasted with institutional regional contexts represented by journals. In addition, a journal's specialty influences SDoH mentions. Specifically, journals on psychiatry, general medicine, and medical specialties tend to mention SDoH more than surgical specialty journals. These findings emphasized the need for a standardized approach to SDoH reporting across varied geographies and specialties.

Notably, our data revealed potential biases in SDoH reporting in the medical literature. Certain SDoH reports, such as sexual orientation with infectious diseases or homelessness with mental disorders, are overemphasized, potentially reinforcing stereotypes or

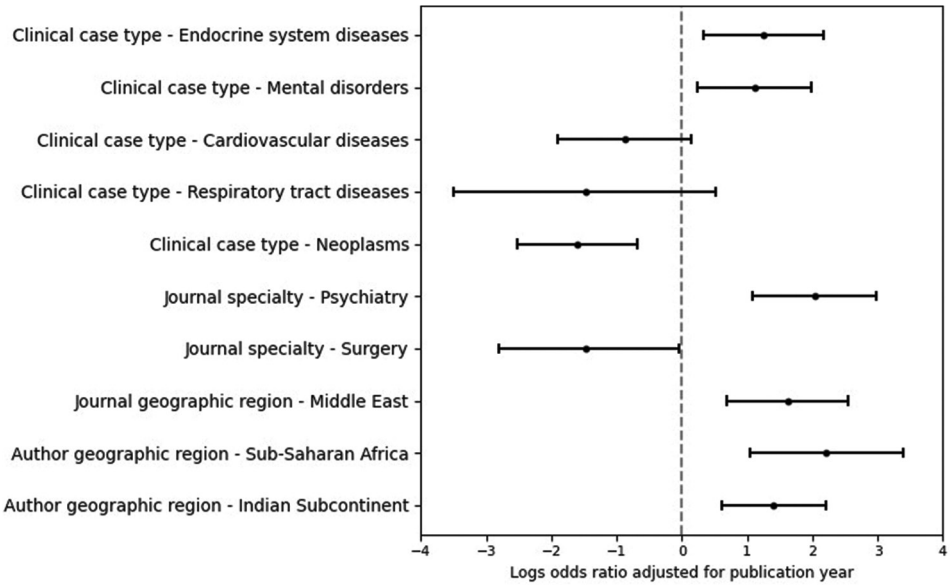

Figure 7. Adjusted odds ratios for the probability of mentioning spiritual beliefs based on clinical case type, journal specialty, journal's geographic region and author's geographic region. The figure was plotted with Matplotlib.

creating oversimplified narratives. Furthermore, these biases risk being duplicated in training large language models, especially those using self-supervised methods with medical literature as data.

4.3. Technological opportunities

Despite the low prevalence of SDoH mentions in clinical case reports, using NER models through Spark NLP offer a potential path for broad-scale clinical record analysis on SDoH mentions. Notably, this method can be used on standard computing hardware,[25] providing access to advanced data analytics. Our research indicated that NER models are more efficient than larger models (e.g., GPT), especially for specific tasks like clinical entity detection. This technology can be used not only for reviewing clinical case reports but also for analyzing EHRs in the search of SDoH,[37,38] thereby enhancing research scalability. In addition, high-level computational analysis could be performed with regular laptops and central processing units (CPUs). Recent studies successfully designed NER models to extract SDoH from clinical narratives.[27] However, the primary objective of our research was not merely to validate these NER models but to analyze the factors associated with the likelihood of mentioning specific SDoH when describing a clinical case.

4.4. Limitations

Our investigation had several limitations that warrant consideration. First, our dataset only included published clinical case reports, which might not reflect the full spectrum of clinical situations or health-care settings. This could lead to a skewed representation of certain regions, affecting our understanding of cultural influences on SDoH mentions.

Second, our analysis might understate SDoH mentions due to two main reasons: our focus was limited to abstracts, specifically sentences outlining primary patient characteristics; and the NER model used had a potential for false negatives, evidenced by the recalls not being 100%. Given the low SDoH mentions in the PubMed corpus, fully evaluating the NER model's recall was challenging. However, our external validation revealed satisfactory recall metrics, and we inferred that the false negatives were likely evenly spread across the model's attribute, subsequently preventing significant impacts on the results from our logistic regression analysis.

In our analysis, we observed that most of the odds ratios (ORs) for the SDoH factors were negative. This finding suggested that specific SDoH mentions within the literature were rare and, when present, were often linked to particular characteristics such as diagnoses, specialties, and cultures. Consequently, this led to OR < 1 for most of the analyzed features. The substantial sample size of our study further amplified the ability of the model to detect statistically significant effects, even for minor associations, adhering to the stringent *p*-value threshold of $P < 0.0001$.

The prevalence of negative ORs could also be due to overadjustment. Overadjustment occurs when a model includes too many variables or inappropriate variables,

leading to biased estimates of the effect size. Despite this risk, the extensive inclusion of variables in our model was a deliberate choice, reflective of the exploratory nature of our research. This project aimed to uncover existing relationships and identify factors potentially associated with SDoH mentions in the literature. To mitigate the risk of arbitrary variable selection, we employed a stepwise approach, including only variables with p-values of 0.001 or less, ensuring that each variable included in the model contributed significantly to the explanatory power of our analysis.

However, we acknowledge that understanding the causality behind these associations requires more sophisticated modeling techniques. Our findings provide the foundation for future research endeavors and in-depth studies that can employ more advanced statistical models to unravel the causal pathways linking SDoH to health outcomes. These studies will be crucial for developing targeted interventions and policies aimed at addressing SDoH more effectively within health-care practices and research.

5. Conclusion

The limited mentions of SDoH in clinical case reports underscore the necessity for better SDoH integration into medical documentation. To mitigate biases in statistical analyses using clinical notes or medical journal content, consistent recording and reporting of SDoH are essential. Spark NLP offers promising avenues for enhancing the extraction and analysis of SDoH from EHRs, highlighting the importance of AI model development to prevent biases that could negatively affect health-care fairness and delivery.

For future research, conducting a similar analysis on the factors associated with SDoH mentions in the full texts of clinical case reports could yield deeper insights. In addition, analyzing actual EHR notes to compare the prevalence and representation of SDoH across different specialties or health-care centers could provide valuable information. Such comparative studies could elucidate the representation and documentation of SDoH across various health-care settings, potentially guiding targeted interventions and policy changes to promote equitable health-care outcomes.

In conclusion, enhancing the documentation and representation of SDoH in the medical literature is critical for advancing toward more informed, equitable, and effective health-care practices and policies. Future studies focused on expanding the scope of analysis to full texts and EHRs could significantly contribute to our understanding and implementation of SDoH in clinical care.

Acknowledgments

None.

Funding

This work has been funded by John Snow Labs Inc.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Julio Bonis, Veysel Kocaman
Formal analysis: Julio Bonis
Investigation: Julio Bonis, Veysel Kocaman
Methodology: Julio Bonis, David Talby
Writing-original draft: Julio Bonis
Writing-review & editing: Veysel Kocaman, David Talby

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The dataset utilized for the logistic regression analysis will be made available upon publication. Interested parties can obtain access for academic purposes by directly contacting the authors (julio@johnsnowlabs.com or veysel@johnsnowlabs.com) and signing a data access agreement.

References

1. McGinnis JM, Williams-Russo P, Knickman JR. The case for more active policy attention to health promotion. *Health Aff (Millwood)*. 2002;21:78-93.

   doi: 10.1377/hlthaff.21.2.78

2. Galea S, Tracy M, Hoggatt KJ, DiMaggio C, Karpati A. Estimated deaths attributable to social factors in the United States. *Am J Public Health*. 2011;101:1456-1465.

   doi: 10.2105/AJPH.2010.300086

3. Hatef E, Kharrazi H, Nelson K, *et al.* The association between neighborhood socioeconomic and housing characteristics with hospitalization: Results of a national study of Veterans. *J Am Board Fam Med.* 2019;32:890-903.

   doi: 10.3122/jabfm.2019.06.190138

4. Hood CM, Gennuso KP, Swain GR, Catlin BB. County health rankings: Relationships between determinant factors and health outcomes. *Am J Prev Med.* 2016;50:129-135.

   doi: 10.1016/j.amepre.2015.08.024

5. Walker RJ, Strom Williams J, Egede LE. Influence of race, ethnicity and social determinants of health on diabetes outcomes. *Am J Med Sci.* 2016;351:366-373.

   doi: 10.1016/j.amjms.2016.01.008

6. Teshale AB, Htun HL, Owen A, *et al.* The role of social determinants of health in cardiovascular diseases: An umbrella review. *J Am Heart Assoc.* 2023;12:e029765.

   doi: 10.1161/JAHA.123.029765

7. Enard KR, Coleman AM, Aver Yakubu R, Butcher BC, Tao D, Hauptman PJ. Influence of social determinants of health on heart failure outcomes: A systematic review. *J Am Heart Assoc.* 2023;12:e026590.

   doi: 10.1161/JAHA.122.026590

8. Ludwig J, Sanbonmatsu L, Gennetian L, *et al.* Neighborhoods, obesity, and diabetes - a randomized social experiment. *N Engl J Med.* 2011;365:1509-1519.

   doi: 10.1056/NEJMsa1103216

9. Wang M, Pantell MS, Gottlieb LM, Adler-Milstein J. Documentation and review of social determinants of health data in the EHR: Measures and associated insights. *J Am Med Inform Assoc.* 2021;28:2608-2616.

   doi: 10.1093/jamia/ocab194

10. Daniel H, Bornstein SS, Kane GC, *et al.* Addressing social determinants to improve patient care and promote health equity: An American college of physicians position paper. *Ann Intern Med.* 2018;168:577-578.

    doi: 10.7326/M17-2441

11. Handerer F, Kinderman P, Tai S. The need for improved coding to document the social determinants of health. *Lancet Psychiatry*. 2021;8:653.

    doi: 10.1016/S2215-0366(21)00208-X

12. Cottrell EK, Dambrun K, Cowburn S, *et al.* Variation in electronic health record documentation of social determinants of health across a national network of community health centers. *Am J Prev Med.* 2019;57:S65-S73.

    doi: 10.1016/j.amepre.2019.07.014

13. Guo Y, Chen Z, Xu K, *et al.* International Classification of Diseases, Tenth Revision, clinical modification social determinants of health codes are poorly used in electronic health records. *Medicine (Baltimore)*. 2020;99:e23818.

    doi: 10.1097/MD.0000000000023818

14. Truong HP, Luke AA, Hammond G, Wadhera RK, Reidhead M, Joynt Maddox KE. Utilization of social determinants of health ICD-10 Z-codes among hospitalized patients in the United States, 2016-2017. *Med Care*. 2020;58:1037-1043.

    doi: 10.1097/MLR.0000000000001418

15. Hatef E, Rouhizadeh M, Tia I, *et al.* Assessing the availability of data on social and behavioral determinants in structured and unstructured electronic health records: A retrospective analysis of a multilevel health care system. *JMIR Med Inform.* 2019;7:e13802.

    doi: 10.2196/13802

16. Guevara M, Chen S, Thomas S, *et al*. Large language models to identify social determinants of health in electronic health records. *NPJ Digit Med*. 2024;7:6.

    doi: 10.1038/s41746-023-00970-0

17. Jiménez Carrillo M, Fernández Rodker J, Sastre Paz M, Alberquilla Menendez-Asenjo Á. ¿Refleja la historia clínica electrónica los determinantes sociales de la salud desde Atención Primaria? [Does the electronic health record reflect the social determinants of health from primary health care?]. *Aten Primaria*. 2021;53:36-42. [In Spanish]

    doi: 10.1016/j.aprim.2020.01.007

18. Gold R, Bunce R, Cowburn S, *et al.* Adoption of social determinants of health EHR tools by community health centers. *Ann Fam Med.* 2018;16:399-407.

    doi: 10.1370/afm.2275

19. Tamang S, Humbert-Droz M, Gianfrancesco M, Izadi Z, Schmajuk G, Yazdany J. Practical considerations for developing clinical natural language processing systems for population health management and measurement. *JMIR Med Inform.* 2023;11:e37805.

    doi: 10.2196/37805

20. Elbattah M, Arnaud É, Gignon M, Dequen G. The Role of Text Analytics in Healthcare: A Review of Recent Developments and Applications: In: *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies* SCITEPRESS - Science and Technology Publications, Vienna, Austria; 2021. p. 825-832.

    doi: 10.5220/0010414508250832

21. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North*. Vol. 1; 2019:4171-4186.

    doi: 10.18653/v1/n19-1423

22. Lee J, Yoon W, Kim S, *et al.* BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36:1234-1240.

    doi: 10.1093/bioinformatics/btz682

23. Haq HU, Kocaman V, Talby D. Mining adverse drug reactions from unstructured mediums at scale. In: *Shaban-Nejad A, Michalowski M, Bianco S, editors. Multimodal AI in Healthcare (Studies in Computational Intelligence)*. Vol. 1060. Cham: Springer; 2023:361-375.

    doi: 10.1007/978-3-031-14771-5_26

24. Kocaman V, Talby D. Accurate clinical and biomedical named entity recognition at scale. *Softw Impacts*. 2022;13:100373.

doi: 10.1016/j.simpa.2022.100373

25. Kocaman V, Talby D. Spark NLP: Natural language understanding at scale. *Softw Impacts*. 2021;8:100058.

doi: 10.1016/j.simpa.2021.100058

26. Zhu Y, Yuan H, Wang S, et al. Large language models for information retrieval: A survey. *arXiv*. Preprint posted online 2023.

doi: 10.48550/arXiv.2308.07107

27. Raza S, Dolatabadi E, Ondrusek N, Rosella L, Schwartz B. Discovering social determinants of health from case reports using natural language processing: Algorithmic development and validation. *BMC Digit Health*. 2023;1:35.

doi: 10.1186/s44247-023-00035-y

28. Hazlehurst B, Naleway A, Mullooly J. Detecting possible vaccine adverse events in clinical notes of the electronic medical record. *Vaccine*. 2009;27:2077-2083.

doi: 10.1016/j.vaccine.2009.01.105

29. Banerji A, Lai KH, Li Y, *et al*. Natural language processing combined with ICD-9-CM codes as a novel method to study the epidemiology of allergic drug reactions. *J Allergy Clin Immunol Pract*. 2020;8:1032-1038.e1.

doi: 10.1016/j.jaip.2019.12.007

30. John Snow Labs Inc. *Detect Sentences in Healthcare Texts*. Available from: https://nlp.johnsnowlabs.com/2021/08/11/sentence_detector_dl_ healthcare_en.html [Last accessed on 2024 Apr 06].

31. John Snow Labs Inc. *Social Determinants of Health*. Available from: https://nlp.johnsnowlabs.com/2023/06/13/ner_sdoh_en.html [Last accessed on 2024 Apr 06].

32. Kocaman V, Talby D. Biomedical named entity recognition at scale. In: Del Bimbo A, Farinella GM, Escalante HJ, *et al*, editors. *Pattern Recognition. ICPR International Workshops and Challenges*. Vol. 12661. Cham: Springer International Publishing; 2021:635-646.

doi: 10.1007/978-3-030-68763-2_48

33. Brown TB, Mann B, Ryder N, *et al*. Language models are few-shot learners. *arXiv*. Preprint posted online 2020.

doi: 10.48550/arXiv.2005.14165

34. OpenAI, Achiam J, Adler S, *et al*. GPT-4 technical report. *arXiv*. Preprint posted online 2023.

doi: 10.48550/arXiv.2303.08774

35. Heinze G, Wallisch C, Dunkler D. Variable selection - A review and recommendations for the practicing statistician. *Biom J*. 2018;60:431-449.

doi: 10.1002/bimj.201700067

36. Tsai J. Lifetime and 1-year prevalence of homelessness in the US population: Results from the national epidemiologic survey on alcohol and related conditions-III. *J Public Health Oxf Engl*. 2018;40:65-74.

doi: 10.1093/pubmed/fdx034

37. Lybarger K, Dobbins NJ, Long R, *et al*. Leveraging natural language processing to augment structured social determinants of health data in the electronic health record. *J Am Med Inform Assoc*. 2023;30:1389-1397.

doi: 10.1093/jamia/ocad073

38. Stewart De Ramirez S, Shallat J, McClure K, Foulger R, Barenblat L. Screening for social determinants of health: Active and passive information retrieval methods. *Popul Health Manag*. 2022;25:781-788.

doi: 10.1089/pop.2022.0228