Journal of the American Medical Informatics Association, 29(1), 2022, 187–196
doi: 10.1093/jamia/ocab199
Advance Access Publication Date: 19 October 2021
Review

Review

The quality of social determinants data in the electronic health record: a systematic review

Lily A. Cook, Jonathan Sachs, and Nicole G. Weiskopf

Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, Oregon, USA

Corresponding Author: Lily A. Cook, PhD Candidate, Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University School of Medicine, Portland, OR, USA; cooli@ohsu.edu

Received 7 July 2021; Revised 24 August 2021; Editorial Decision 30 August 2021; Accepted 8 September 2021

ABSTRACT

Objective: The aim of this study was to collect and synthesize evidence regarding data quality problems encountered when working with variables related to social determinants of health (SDoH).
Materials and Methods: We conducted a systematic review of the literature on social determinants research and data quality and then iteratively identified themes in the literature using a content analysis process.
Results: The most commonly represented quality issue associated with SDoH data is plausibility ($n = 31$, 41%). Factors related to race and ethnicity have the largest body of literature ($n = 40$, 53%). The first theme, noted in 62% ($n = 47$) of articles, is that bias or validity issues often result from data quality problems. The most frequently identified validity issue is misclassification bias ($n = 23$, 30%). The second theme is that many of the articles suggest methods for mitigating the issues resulting from poor social determinants data quality. We grouped these into 5 suggestions: avoid complete case analysis, impute data, rely on multiple sources, use validated software tools, and select addresses thoughtfully.
Discussion: The type of data quality problem varies depending on the variable, and each problem is associated with particular forms of analytical error. Problems encountered with the quality of SDoH data are rarely distributed randomly. Data from Hispanic patients are more prone to issues with plausibility and misclassification than data from other racial/ethnic groups.
Conclusion: Consideration of data quality and evidence-based quality improvement methods may help prevent bias and improve the validity of research conducted with SDoH data.

Key words: data quality, social determinants of health, healthy equity, bias, Hispanic Americans

INTRODUCTION

Interest in social determinants of health (SDoH) among clinicians, researchers, and policy-makers has increased in recent years, driven both by a recognition of their role as major contributors to health outcomes and by interest in improving health equity. There are substantial and justifiable concerns, however, regarding the quality of SDoH in clinical data. It is a long-established tenet of information science that poor-quality data lead to poor-quality results.[1] Without attention to the quality of SDoH data, researchers cannot guarantee that results provide valid or useful insights.

Our objective was to conduct a review of the literature on SDoH data quality to characterize the issues that impact the use of these data for research and policy. Specifically, our goal was to collect and synthesize available evidence regarding the kinds of quality issues typically encountered with SDoH data, the biases these issues may create during analysis, and any identified methodological solu-

tions to these issues that can be used by researchers to improve social determinants data quality prior to analysis.

We were unable to find any prior work that both gathers information about how researchers can improve the quality of clinical SDoH data and also summarizes how specific issues of bias and validity are introduced into research utilizing social determinants variables. To the best of our knowledge, this review is the first to bring together information about a variety of social determinants variables to examine the issues inherent to the field of social informatics.

BACKGROUND

Although data quality—also known as data integrity, data accuracy, or data validation—was initially conceptualized for broad application across information systems regardless of context, there is a growing body of literature devoted to the topic of electronic health record (EHR) data quality. As the field of data science has shifted from collecting data toward processing the massive amounts of data already collected, informatics researchers are addressing the problem of secondary use—how to make clinical documentation usable as a research source.

Published in 2016, the Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data[2] provides 3 major concepts that describe the quality of EHR data used for research:

1. *Conformance* speaks to whether the dataset's reported values meet structural standards and formats. Conformance is further broken into 3 subcategories: *value conformance*, *relational conformance*, and *computational conformance*.
2. *Completeness* looks at whether or not the data are actually present.
3. *Plausibility* asks if the data values are believable and accurate. It can be broken into 3 additional subcategories: *uniqueness plausibility*, *atemporal plausibility*, and *temporal plausibility*.

SDoH are, by their very nature, nonmedical. As such, few SDoH data elements are routinely stored in the medical record. Some, such as race and health insurance status, are commonly collected during patient onboarding and are usually available as demographic or administrative data. However, these elements represent only a small portion of all the nonclinical factors known to influence health, and although there has been a widespread effort to collect more SDoH from patients these data have remained sparse.[3] Although an exhaustive list of all methods used by researchers to access a wider array of social determinants factors is beyond the scope of this work, researchers have generally relied on the following:

1. *Diagnostic codes (eg, ICD Z-codes)* indicating SDoH such as "Problems relating to housing and economic circumstances" have the benefit of being standardized across systems, but in practice are rarely used by clinicians.[4–6]
2. *Geocoding patient addresses* allow researchers to integrate biomedical data from the EHR with community-level data sources such as the US Census. In 2014, the Institute of Medicine suggested using "neighborhood and community composition" as a proxy for individual-level indicators that cannot be directly collected from patients.[7]
3. *Structured and semistructured tools for clinicians* such as flowsheets, screening tools, and questionnaires, can collect more information from patients than can be found in administrative fields. Although a variety of these tools are now available, including the Protocol for Responding to and Assessing Patients' Assets, Risks, and Experiences (PRAPARE) and the Epic SDoH Wheel, they have been adopted by a very small number of clinics. The lack of any single, ubiquitously applied clinical tool means extra work for researchers, who would need to extract and harmonize any information gathered from these applications.

Each of these methods has benefits and drawbacks that must be considered when selecting a research dataset. However, little information is available to assist researchers making these choices, and there are no agreed-upon best practices for working with SDoH data.

MATERIALS AND METHODS

Search strategy and screening

We conducted an iterative, deductive, and systematic literature review. Specifics of the workflow are detailed in Figure 1. First, PubMed and Ovid MEDLINE databases were searched for articles focused on social determinants research and data quality, accuracy, validity, or the introduction of bias. Although no date range was specified for articles, the initial results were limited by the date range of databases themselves. The search terms were constructed to align with the definition of SDoH created by the World Health Organization (WHO) and utilized by

major organizations such as Gravity Project.[8,9] Nonmodifiable social and economic factors such as race/ethnicity, socioeconomic status, education level, environmental health (proximity to healthy food, walkability, exposure to environmental toxins, etc.), and health insurance status were all considered SDoH for the purpose of this review. However, in keeping with the WHO's definition, modifiable behaviors such as smoking and exercise were not. Because data linkage is commonly used to enrich clinical social determinants datasets with community-level information, articles about the quality of geocoded patient address data were included if they discussed research focused on linking clinical data to exterior datasets for clinical research purposes.

The Medical Subject Headings (MeSH) database was used to identify appropriate keywords. The initial search was conducted in PubMed, and an adjacency search was performed in Ovid MEDLINE to find articles not indexed in PubMed. Details of the search strategy, including keywords, can be found in Supplementary Appendix SA.

Eligibility

The results of the 2 database searches were compared in order to identify and remove duplicate articles, and then the remaining articles were screened based on title and abstract. After screening, the first author then manually reviewed the remaining articles to determine whether they met the eligibility criteria described below and summarized in Table 1.

Inclusion criteria

Articles were eligible for the review if they were in English, used data from healthcare systems in the United States and Canada, and were original research. The research described in the articles must use patient health data originating from clinical sources (ie, data sourced from electronic health records, disease registries, etc.). Studies examining information from registries were included because this

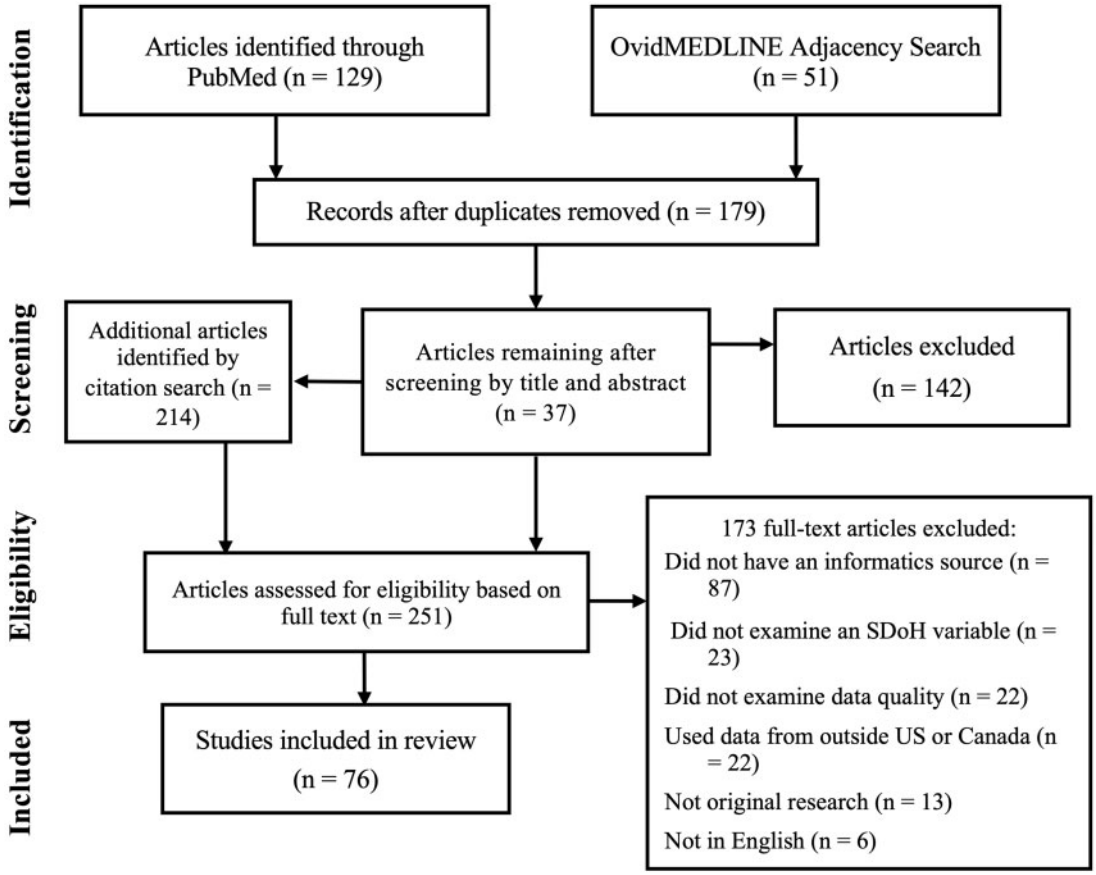

Figure 1. PRISMA flow diagram.

Table 1. Eligibility criteria for articles

	√ Included:	X Excluded:
Topic/Focus	Original, peer-reviewed research focused on the quality of social determinants of health data.	Reviews; opinion pieces; research that has not been peer-reviewed.
Social Determinants of Health Factors	Race/ethnicity, language preference, health insurance status, country of origin, occupation, socioeconomic status, education level, environmental health (proximity to healthy food, walkability, exposure to environmental toxins, etc.), geocoded patient address data (only included if the article primarily focused on linking clinical data to external datasets for research on social determinants)	Behaviors (eg, smoking and exercise)
Sources of Health Data	Clinical sources within the United States and Canada: EHR, medical registries, administrative databases compiled from EHR data, observational studies using clinical data pulled directly from the medical records of participants	Nonclinical sources: population-level data, mHealth sources, genomic datasets, vital records (ie, birth or death certificate data); Clinical sources outside the United States or Canada
Language	Articles written in English.	Articles in languages other than English.

information is often abstracted directly from medical records.[10] Also included were several large databases compiled from electronic health record EHR data, such as the Biomedical Translational Research Information System repository,[11] and the Healthcare Cost and Utilization Project's State Inpatient Database.[4,12–14] Research using datasets from large cohort studies such as the National Birth Defect Prevention Study were included if the study's dataset was drawn directly from the participant's medical records.[15,16]

Exclusion criteria

Articles were ineligible for this review if they described research that used patient health data from nonclinical sources (eg, population-level data, patient-generated data such as mHealth sources, or data from genomic datasets not originating from clinical sources). Vital records such as birth and death certificate data were not included because they often differ significantly in structure and content from other medical records.

To identify additional works missed by the initial query, a snowball technique was applied to the citations in the eligible articles.

Data extraction and thematic analysis

Using a deductive approach, content analysis was performed on eligible articles. Each manuscript was categorized by (1) the specific social determinant examined and (2) the primary data quality issue.[17] This information was then abstracted and tabulated.[18] Data abstraction was performed alongside a closer reading of the selected articles, which informed the thematic analysis.[19] An iterative, inductive approach was taken to identify themes in the literature, with a specific focus on themes that could be actionable to researchers using health records for social determinants research.[20] Once themes were identified, the articles were reviewed a final time to abstract and tabulate the prevalence of issues and common approaches to solutions. A complete list of the categories selected for each article can be found in Supplementary Appendix SB.

RESULTS

A total of 76 articles were included in this review. Throughout the literature, the most commonly represented quality issue associated with social determinants data was *plausibility*, that is, accuracy ($n = 31$, 41%). Thirty-eight percent ($n = 29$) of manuscripts focused primarily on the *completeness* of social determinants data in the medical record—whether or not data were missing. The remaining 21% looked largely at *conformance*—whether data were compatible ($n = 16$). A tabulated breakdown by data quality issue and social determinants category is available in Table 2.

Articles about race, ethnicity, country of origin, and language preference were grouped into a single category and had the largest body of data quality literature (40 of 76 articles; 53%). Sixteen

articles (21%) addressed the quality of geocoded patient address data, which is frequently used in

clinical social determinants research to link individual, patient-level data to community-level datasets to incorporate variables not available in the medical record. Also, represented in the literature were occupation (9%), environmental factors (7%), and insurance status (1%). Seven articles (9%)

addressed social determinants data generally without focusing on any specific variable. Three of these nonspecific articles discussed the use of International Classification of Diseases (ICD) codes (a.k.a., Z-codes) as a source of social determinants information in the medical record.

Bias

The first theme identified in the thematic analysis was that bias or validity problems were likely to result from data quality concerns. A majority of articles (47 of 76; 62%) either found bias when running test analyses on their datasets or they noted that data quality was differentially poor for certain groups and thus there was a high potential for bias. Twenty-four articles (32%) did not evaluate their datasets for bias and 5 articles (7%) tested for bias but were unable to find any. Sixteen articles (21%) observed that data are differentially incomplete (also referred to as Missing Not at Random); 23 (30%) noted misclassification bias. Results about bias associated with specific social determinants are presented below and summarized in Table 3.

Race/ethnicity/country-of-origin variables

Plausibility

Articles that discussed race, ethnicity, or country-of-origin data were most concerned with

plausibility (ie, accuracy), which was discussed in 26 of the 40 articles (65%). Eighty-five percent of these articles (22 of 26) noted the potential for implausible data to cause error or bias in research, most commonly with misclassification. Three studies (12%) looked for bias but did not find any.[20–22]

Misclassification bias, that is, incorrect assignment, was noted as a problem or a potential problem in 18 of the 26 (69%) articles about the plausibility of race/ethnicity data.[11,13,23–38] Further, several studies reported that implausible data and misclassification errors were more likely for certain groups:

- Fourteen studies reported that Hispanic patients were more likely to be misclassified, either that information about their ethnicity was missing or they had been mistakenly grouped into the "Other" category.[11,23–25,29,30,32,33,36–41]
- Four of these studies also found that Asian patients were more likely to be missing information identifying their race than white patients.[24,30,33,36]
- Six studies found disproportionately high rates of misclassification for American Indians in comparison to other racial/ethnic groups; most often, these patients were misidentified as white.[27,28,30,32,41,42]

Completeness

The remaining 15 studies that looked at race/ethnicity data primarily examined its *completeness* (38%). Ten of these (66%) identified that the incomplete data led to validity issues, most commonly that these data were not missing at random and had the potential to introduced bias.[12,34,43–50]

Geocoded patient address data used for linkage to community-level variables

Articles about geocoded patient address data, on the other hand, were largely concerned with *relational conformance* (ie, linkage match rates for geocoding), which was examined in 10 of the 16 articles (63%). Plausibility, that is accuracy, was the primary concern of 4 of the articles about geocoded patient address data.

Overall, 43% ($n = 7$) of the studies about geocoded patient address data acknowledged or established a potential for bias in their datasets. The remainder looked at data quality but did not evaluate their datasets for any validity problems that may result. Although race/ethnicity data were mostly plagued with a single type of error (ie, misclassification), geocoded address data linked with community-level data were associated with multiple forms, including cartographic confounding and Type II error (ie, falsely accept the null hypothesis).[51]

Another study characterized the issues encountered with geocoded patient data as the distinction between individual- and community-level variables. They found that "the accuracy of the community-level data for identifying patients with and without social risks was 48.0%." The authors noted that the use of these data for patient risk stratification "may heighten the risk of ecologic fallacy, wherein incorrect assumptions are made about an individual based on aggregate-level information about a group."[52]

It should also be noted that the quality problems found with geocoded patient address data were not randomly distributed; for ex-

Table 2. Characteristics of studies included in this review

	Primary social determinant of health, *n* (%)					
Primary data quality issue	Race, ethnicity, country of origin	Insurance status	Occupation	General community-level	Environmental	Nonspecific
Completeness (missing data), *n* = 29	15 (37.5%)	1 (100%)	6 (86%)	2 (12.5%)	1 (20%)	6 (86%)
Conformance (incompatible data), *n* = 16	0	0	1 (14%)	10 (62.5%)	3 (60%)	0
Plausibility (inaccurate data), *n* = 31	25 (62.5%)	0	0	4 (25%)	1 (20%)	1 (14%)
Total, *n* = 76	**40**	**1**	**7**	**16**	**5**	**7**
Typical article title	*"Accuracy of Race, Ethnicity, and Language Preference in an Electronic Health Record"*	*"Primary Payer at DX: Issues with Collection and Assessment of Data Quality"*	*"Availability and accuracy of occupation in cancer registry data among Florida firefighters"*	*"Match Rate and Positional Accuracy of Two Geocoding Methods for Epidemiologic Research"*	*"Residential mobility in early childhood and the impact on misclassification in pesticide exposures"*	*"Utilization of Social Determinants of Health ICD-10 Z-Codes Among Hospitalized Patients in the United States, 2016–2017"*
Usual source for this information within the patient record	administrative or demographic sources			patient address is geocoded to link community-level data		diagnosis codes

Table 3. Findings about bias and differential data quality

Bias Finding?	Social determinant	Bias type	Articles reporting that finding, *n* (%)
Yes	Race/ethnicity	Misclassification	19 (25.0)
		Missing Not at Random (MNAR)	9 (11.8)
		Differentially implausible	2 (2.6)
		Other	1 (1.3)
	Insurance	Missing Not at Random (MNAR)	1 (1.3)
	Occupation	Missing Not at Random (MNAR)	4 (5.3)
	General Community Level	Rural data are problematic	3 (3.9)
		Other	3 (3.9)
	Environmental	Misclassification	3 (3.9)
	Nonspecific	Missing Not at Random (MNAR)	2 (2.6)
Unknown	Did not evaluate for bias		24 (31.6)
No	Evaluated for bias and found none		5 (6.6)

ample, relational conformance (ie, match rates) tend to be poorer for rural areas and certain parts of the country.[53–55]

Environmental health variables

Because information about exposure to toxins is rarely recorded in the medical record, geocoded patient addresses are used to link health information to data about the environment. We found 5 studies discussing the use of patient addresses for exposure assessment. As with the geocoded community-level variables discussed above, *relational conformance* was represented in the majority of the articles about the quality of the datasets used for environmental health (60%).

Four (80%) of the included articles explored the impact of residential mobility on exposure assessment; that is, whether patients' moving impacted the results of studies looking at environmental outcomes.[15,16,56,57] In all of these articles, bias was characterized as misclassification of exposure to contaminants, an issue that has been noted elsewhere to be a source of Type II error in environmental health research.[58]

Nonspecific social determinants

Six of the 7 articles which addressed social determinants data as a broad, general category assessed the *completeness* of this information (86%); one addressed *plausibility*. Two of the articles mentioned the potential for bias, in both cases due to data missing nonrandomly.[6,59] Three of the articles looked at the completeness of ICD Z- or V-codes, diagnostic codes that can be used by clinicians to collect SDoH data from patients.[4–6] All 3 articles concluded that clinicians were utilizing ICD codes to represent SDoH around 2% of the time.

Occupation

Six of the 7 studies (86%) that looked at the quality of occupational data primarily examined completeness, all noting that occupational information is frequently missing from the health record.[10,60–64] Four studies found that data were not missing at random and that male patients were more likely to have occupational information in their record.[60,62–64]

Recommendations from the literature

The second theme we identified in our analysis is that there are solutions researchers can use to mitigate the issues caused by data quality problems. Forty-seven of the articles (62%) made at least one evidence-based recommendation for researchers seeking to improve the quality of social determinants data after it has been collected. We grouped these recommendations into 5 suggestions, which are detailed below and briefly summarized in Table 4.

Table 4. Summary of recommendations found in the articles

Five ways to increase data quality	
Recommendation	References supporting this recommendation
1. Avoid complete case analysis	23,50,65
2. Impute data	14,22,29,30,36,37,45,46,50,61,65–71
3. Rely on multiple sources	13,26,28,49,60,62,64,72
4. Use validated software tools	10,12,54,60,62,73–77
5. Select addresses thoughtfully	15,16,56,57

Avoid complete case analysis

It is a common practice to exclude incomplete (ie, missing) data from the analysis, a method also known as casewise deletion or complete case analysis. However, 3 studies in our review found that casewise deletion decreased the quality of race/ethnicity data.[23,50,65] Grundmeier et al[65] found that

using only complete cases "produced highly biased results," and in fact reversed the odds ratio for the Black subjects in their dataset. Brown et al[50] found that their "race and ethnicity coefficient estimates are often biased downwards either toward zero or more negative when data with missing race and ethnicity is dropped." In all 3 studies, imputation was recommended as preferable to casewise deletion of missing data.

In a study on using patient addresses to determine pesticide exposures, Ling et al[57] noticed that there were significant differences between patients with complete address information and those who were missing information. In particular, Hispanic women born in Mexico and people living in poor neighborhoods were more likely to have missing addresses. For studies that rely on patient address to determine exposure, this means that these groups are more likely to be excluded from the analysis, potentially biasing the results.

One additional study about geocoding of patient addresses did not evaluate for bias, but did note that "unmatched addresses tend to be unevenly distributed—more likely to occur in rural areas and newly developed suburban areas, and less likely to occur in inner-city areas."[53] In other words, complete case analysis would likely exclude a disproportionate number of rural patients.

Impute data

Several studies looked at the use of imputation, also referred to as indirect estimation, to increase the completeness of datasets and avoid casewise deletion.[14,22,30,45,46,50,61,65,67–70] Imputation is a way to infer missing data, and there are many imputation methods that can be used to generating substitute values to fill in missing data. Most of these studies examined methods for imputing race/ethnicity data, although several looked at imputing geocoded patient addresses,[67,69,70] and one looked at imputing occupational data.[61]

The most widely researched imputation method was Bayesian Improved Surname Geocoding (BISG).[22,45,46,50,65] BISG is used to supplement missing race/ethnicity data and provides a probability of a patient belonging to a particular racial or ethnic group based on that patient's geocoded address and their last name. In their study on the use of BISG, Dembosky et al[46] found that it "did not substantially alter the estimated overall racial/ethnic distribution, but it did modestly increase sample size and statistical power."

Imputation was recommended not only as a solution for missing data but also to improve the accuracy of implausible data.[29,30,36,37,66,71] Methods involving Spanish surname coding, a close relative of BISG that uses a patient's last name to guess their ethnicity, were investigated in several studies to increase the reliability and consistency of data from Hispanic patients.[29,36,37,66] Two articles validated similar, surname-based methods for data from Asian/Pacific Islander patients.[30,71] All of these studies confirmed that these techniques reduced misclassification errors and enhanced data quality for their respective populations.

Finally, 2 additional studies looked at imputing race/ethnicity data with anonymized clinical datasets,[14,68] a situation where patient identifying information has been removed, and therefore, it is not possible to use imputation methods that rely on patient surname or geocoded address. Ma et al[14] compared 4 imputation methods and found that conditional multiple imputation, "substantially improved statistical inferences for racial health disparities research."

Rely on multiple sources

This method compares and links data across an individual's record in order to fill in or correct missing data fields and was recommended by several studies as a way to either increase completeness[26,28,49,60,62,64,72] and/or check the accuracy of implausible data.[13,26,28] For example, Smith et al.[26] recommended supplementing race/ethnicity data from the EHR with birth certificate data to increase completeness and plausibility. Another study used natural language processing (NLP) to "improve the identification of race and ethnicity in EHR data."[49] Those researchers used NLP to comb through the unstructured text fields in clinical notes, then used the results to augment race/ethnicity data missing from structured fields.

Use validated software tools

Several articles evaluated a specific software tool, usually one developed by the authors, for either assessing or increasing the quality of the data. Articles about geocoding patient address data, particularly, evaluated the ability of specific geocoding tools, such as ArcGIS, to increase data quality by improving match rates or positional accuracy.[54,73–75] For researchers seeking to increase the quality of occupation data prior to analysis, the National Institute of Occupational Safety and Health (NIOSH) has created a free, web-based system called NIOSH Industry and Occupation Computerized Coding System[78] that was evaluated and recommended by several studies.[10,60,62,76] Some studies tested data quality assessment tools such as the Data Quality Assessment Tool, created to evaluate patient records at Community Health Centers, and the Data Completeness Analysis Package.[12,77]

Select addresses thoughtfully

Patients move over addresses over time, which means that researchers often have decisions to make about which address to use. When confronted with data from patients who may have moved several times over the study period, 3 studies concluded that a single patient address was sufficient;[15,56,57] one study recommended using the most recent address,[56] while the other 2 concluded that address at birth was adequate for research on the effect of early exposures.[15,57] However, a study by Brokamp et al[16] compared the effect of using the most recent patient address, birth address, and an average across various addresses. They found that using the most recent address or address at birth for a mostly urban population monitored over a 7-year period could create a bias toward the null.

DISCUSSION

The data quality issues represented in our review found in similar proportions to those in Weiskopf and Weng's[79] review examining the quality of EHR data for secondary use. These both validate our findings and suggest that SDoH data suffer from quality problems similar to other data in the medical record.

Our review found that the category of data quality problem varies depending on the variable. Likewise, the kind of error created by these data quality problems also varies based on the social determinant factor in question. Most notably, problems encountered with the quality of SDoH data do not occur randomly. Although many researchers are aware that data "missing not at random," commonly abbreviated to MNAR, can cause bias during analysis, fewer are aware of the problems associated with other kinds of "data quality not at random." However, problems with plausibility not at random—for example, the accuracy of data for Hispanic or Latino patients being lower than the accuracy of data from white patients in the same dataset—has similarly profound implications. Namely, when patients from one racial or ethnic group are lost in another group or mistakenly categorized as "Other," subsequent analysis can cause those groups to be under-represented in research results. Misidentification of the race or ethnicity of groups of patients can inadvertently lead to the erasure of those groups from clinical research.

Several articles documented that race/ethnicity/country-of-origin data tend to be recorded inconsistently across a patient's record, especially for Hispanic patients. Why is data quality so poor for this group? Thirteen studies speculated that this may be due to the "fluid, debatable, and problematic"[34] nature of the definition of race and ethnicity.[11,22,24–26,34,37,40,41,45,46,80,81] Race is, after all, not a biological category but a social and political one,[82] and thus its terminology shifts over time. Pellegrin et al[40] noted that the fluidity of these definitions leads patients to respond inconsistently to questions about their race/ethnicity, thus causing problems with data *reliability*. Further, the fact that these categories are so broad and poorly defined leads to difficulties with data *validity*.

At the institutional level, several studies speculate that the quality of data about ethnicity has been impaired by variations in how healthcare systems record and handle this information. As one study noted, "inconsistent classification of Hispanics is likely attributable, in part, to differences in the definition of being Hispanic across clas-

sification schemes."[27] Because the gold standard for race/ethnicity is widely considered to be patient self-report,[13,24,32,34,37,46,80,83] it is possible that the increasing use of dynamic patient-facing data entry tools may allow people to inform and correct their own demographic information, thus helping to improve the quality of race, ethnicity, and country-of-origin data in the future.[80]

The quality of patient elements, particularly demographic data, promises to become increasingly important as the efforts to link patient records across multiple institutions are expanded. This is necessary for large-scale research, big data analytics, and continuity of patient care. Privacy-preserving record linkage (PPRL) methods identify when records from different sources belong to the same entity while minimizing the exposure of sensitive personal information. These techniques often rely heavily on patient address along with name and date of birth. When there are errors or missing address data, linkage quality suffers.[84]

Many of the social determinants data elements most commonly used in research, such as race, ethnicity, insurance status, and address, were originally collected as demographic data for administrative purposes. Inevitably, data quality issues arise when these elements are used for secondary, retrospective research.[85] Given the increasing importance of social determinants in health equity research and intervention, it is crucial that healthcare institutions work to improve the quality and availability of these data. Efforts such as the Gravity Project are already underway to create standardized, structured reporting of SDoH. Consistently applied standards for SDoH data collection in the EHR would result in improved data quality, which in turn would lead to more robust research, care coordination, and population health management.

Limitations

The data quality concepts used here are not completely orthogonal or distinct. For example, several studies found that the plausibility of the patient race/ethnicity information in their dataset was questionable because the data were incomplete; in other words, plausibility was low because a patient's race was reported correctly in one area, but as "other" in another.[21,86,87]

The snowball sampling approach we used may have caused some research areas to be under-represented. Citation searching is inherently exponential; if our initial search turned up few articles within a certain domain, then that domain may appear to have a smaller body of literature than is in fact present. In addition, publication bias may have affected our findings if authors did not report negative findings when evaluating their datasets for bias.

Although some social determinants variables may have more thoroughly documented data quality issues, this does not mean that those variables are of poorer quality. A larger body of research may indicate simply that these variables are more accessible to researchers and therefore easier to study. For example, our finding that articles about race/ethnicity data were such a large proportion of the literature may reflect that this information is more readily available in the structured fields of the health record than other SDoH variables.

Another limitation is that our search for solutions to these data quality problems is focused on the needs of researchers using observational databases. Because researchers require ex post facto methods for improving secondary use data, we ignored any recommendations from in the literature about improving data collection practices at the point of care.

CONCLUSIONS

The types of quality problems found with SDoH data vary depending on the variable; race/ethnicity data from the health record can be implausible or incomplete, while linked community-level data are prone to problems with nonconformance as well as plausibility. Similarly, data quality problems can lead to corresponding issues of validity and reliability; race/ethnicity data that are implausible or missing not at random may lead to misclassification bias, while problems with geocoding can lead to misclassification, confounding, or ecologic fallacy. Several studies have documented that data quality from Hispanic patients can be particularly implausible and is especially prone to misclassification bias when compared to data from other racial/ethnic groups.

Fortunately, evidence-based solutions are available for researchers who want to improve the quality of social determinants data ex post facto. While complete case analysis has the potential for bias, imputation techniques can avoid these shortcomings. Consideration of data quality by researchers prior to analysis, along with thoughtfully applied quality improvement methods, may help prevent bias and improve the validity of research conducted with SDoH data.

FUNDING

LAC has received support from the National Library of Medicine under Award Number T15LM007088. NGW has received funding under the National Library of Medicine Award Numbers K01LM012738 and R21LM013645.

AUTHOR CONTRIBUTIONS

LAC led the study, drafted the manuscript, conducted the literature review, and analyzed the data. JS helped revise the manuscript and drafted significant portions of the "Discussion" section. NGW assisted with the interpretation of the data, contributed to the conception and design of the study and to the structure of the manuscript, and also provided substantial intellectual content and critical revision.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

The authors would like to thank Andrew S. Hamilton for his invaluable assistance with the literature search and Michael Kahn for his helpful feedback and guidance.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

The data underlying this article are available in the online supplementary material (see Supplementary Appendix SB).

REFERENCES

1. Work with new electronic 'brains' opens field for Army math experts. *The Hammond Times*. 1957 Nov 10: 65.
2. Kahn MG, Callahan TJ, Barnard J, *et al.* A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS (Wash DC)* 2016; 4 (1): 1244.
3. 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 (3): e13802.
4. Torres JM, Lawlor J, Colvin JD, *et al.* ICD social codes: an underutilized resource for tracking social needs. *Med Care* 2017; 55 (9): 810–6.
5. 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 (52): e23818.
6. 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 (12): 1037–43.
7. IOM (Institute of Medicine). *Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2*. Washington, DC: The National Academies Press; 2014.
8. World Health Organization (WHO). Social determinants of health 2021. https://www.who.int/health-topics/social-determinants-of-health#tab=tab_1 Accessed June 15, 2021.
9. Arons A, DeSilvey S, Fichtenberg C, Gottlieb L. Documenting social determinants of health-related clinical activities using standardized medical vocabularies. *JAMIA Open* 2019; 2 (1): 81–8.
10. Freeman MB, Pollack LA, Rees JR, *et al.*; Enhancement of NPCR for Comparative Effectiveness Research Team. Capture and coding of industry and occupation measures: findings from eight National Program of Cancer Registries states. *Am J Ind Med* 2017; 60 (8): 689–95.
11. Magana Lopez M, Bevans M, Wehrlen L, Yang L, Wallen GR. Discrepancies in race and ethnicity documentation: a potential barrier in identifying racial and ethnic disparities. *J Racial Ethn Health Disparities* 2016; 4: 812–8.
12. Nasir A, Liu X, Gurupur V, Qureshi Z. Disparities in patient record completeness with respect to the health care utilization project. *Health Informatics J* 2019; 25 (2): 401–16.
13. Zingmond DS, Parikh P, Louie R, *et al.* Improving hospital reporting of patient race and ethnicity—approaches to data auditing. *Health Serv Res* 2015; 50 (Suppl 1): 1372–89.
14. Ma Y, Zhang W, Lyman S, Huang Y. The HCUP SID imputation project: improving statistical inferences for health disparities research by imputing missing race data. *Health Serv Res* 2018; 53 (3): 1870–89.
15. Chen L, Bell EM, Caton AR, Druschel CM, Lin S. Residential mobility during pregnancy and the potential for ambient air pollution exposure misclassification. *Environ Res* 2010; 110 (2): 162–8.
16. Brokamp C, LeMasters GK, Ryan PH. Residential mobility impacts exposure assessment and community socioeconomic characteristics in longitudinal epidemiology studies. *J Expo Sci Environ Epidemiol* 2016; 26 (4): 428–34.
17. Bryman A. *Social Research Methods*. 5th ed. New York, NY: Oxford University Press; 2016: 314 p.
18. Huberman AM, Miles MB. *Qualitative Data Analysis: An Expanded Sourcebook*. 2nd ed. Thousand Oaks, CA: Sage Publications; 1994.
19. Ezzy D, Liamputtong P. *Qualitative Research Methods: A Health Focus*. South Melbourne, VIC: Oxford University Press; 1999.
20. Boyatzis RE. *Transforming Qualitative Information: Thematic Analysis and Code Development*. Thousand Oaks, CA: Sage Publications; 1998.
21. Hamilton NS, Edelman D, Weinberger M, Jackson GL. Concordance between self-reported race ethnicity and that recorded in a veteran affairs electronic medical record. *N C Med J* 2009; 70 (4): 296–300.
22. Derose SF, Contreras R, Coleman KJ, Koebnick C, Jacobsen SJ. Race and ethnicity data quality and imputation using U.S. Census Data in an integrated health system: the Kaiser Permanente Southern California Experience. *Med Care Res Rev* 2013; 70 (3): 330–45.
23. Lee SJ, Grobe JE, Tiro JA. Assessing race and ethnicity data quality across cancer registries and EMRs in two hospitals. *J Am Med Inform Assoc* 2016; 23 (3): 627–34.
24. Grafova IB, Jarrin OF. Beyond black and white: mapping misclassification of Medicare beneficiaries race and ethnicity. *Med Care Res Rev* 2021; 78 (5): 616–26.
25. Webster PS, Fulton JP, Sampangi S. Conflicting race/ethnicity reports: lessons for improvement in data quality. *J Registry Manag* 2013; 40 (3): 122–6.
26. Smith N, Iyer RL, Langer-Gould A, *et al.* Health plan administrative records versus birth certificate records: quality of race and ethnicity information in children. *BMC Health Serv Res* 2010; 10: 316.
27. Gomez SL, Glaser SL. Misclassification of race/ethnicity in a population-based cancer registry (United States). *Cancer Causes Control* 2006; 17 (6): 771–81.
28. Bigback KM, Hoopes M, Dankovchik J, *et al.* Using record linkage to improve race data quality for American Indians and Alaska Natives in two Pacific Northwest State Hospital discharge databases. *Health Serv Res* 2015; 50 (Suppl 1): 1390–402.
29. Pinheiro PS, Sherman R, Fleming LE, *et al.* Validation of ethnicity in cancer data: which Hispanics are we misclassifying? *J Registry Manag* 2009; 36 (2): 42–6.
30. Eicheldinger C, Bonito A. More accurate racial and ethnic codes for Medicare administrative data. *Health Care Financ Rev* 2008; 29 (3): 27–42.
31. Fiscella K, Fremont AM. Use of geocoding and surname analysis to estimate race and ethnicity. *Health Serv Res* 2006; 41 (4 Pt 1): 1482–500.
32. West CN, Geiger AM, Greene SM, *et al.* Race and ethnicity: comparing medical records to self-reports. *J Natl Cancer Inst Monogr* 2005; (35): 72–4.
33. Waldo DR. Accuracy and bias of race ethnicity codes in the Medicare enrollment database. *Health Care Financ Rev* 2004; 26 (2): 61–72.
34. 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.
35. Swallen KC, Glaser SL, Stewart SL, West DW, Jenkins CNH, McPhee SJ. Accuracy of racial classification of Vietnamese patients in a population-based cancer registry. *Ethn Dis* 1998; 8 (2): 218–27.
36. Polednak AP. Agreement in race-ethnicity coding between a hospital discharge database and another database. *Ethn Dis* 2001; 11 (1): 24–9.
37. Morgan RO, Wei II, Virnig BA. Improving identification of Hispanic males in Medicare: use of surname matching. *Med Care* 2004; 42 (8): 810–6.
38. Lauderdale D, Goldberg J. The expanded racial and ethnic codes in the Medicare data files: their completeness of coverage and accuracy. *Am J Public Health* 1996; 86 (5): 712–6.
39. Lee W-C, Veeranki SP, Serag H, Eschbach K, Smith KD. Improving the collection of race, ethnicity, and language data to reduce healthcare disparities: a case study from an Academic Medical Center. *Perspect Health Inf Manag* 2016; 13 (Fall): 1g.
40. Pellegrin KL, Miyamura JB, Ma C, Taniguchi R. Improving accuracy and relevance of race/ethnicity data: results of a statewide collaboration in Hawaii. *J Healthc Qual* 2016; 38 (5): 314–21.
41. Gomez SL, Kelsey JL, Glaser SL, Lee MM, Sidney S. Inconsistencies between self-reported ethnicity and ethnicity recorded in a health maintenance organization. *Ann Epidemiol* 2005; 15 (1): 71–9.
42. Fiscella K, Meldrum S. Race and ethnicity coding agreement between hospitals and between hospital and death data. *Med Sci Monit* 2008; 14 (3): SR9–13.
43. Gomez SL, Glaser SL, Kelsey JL, Lee MM. Bias in completeness of birthplace data for Asian groups in a population-based cancer registry (United States). *Cancer Causes Control* 2004; 15 (3): 243–53.
44. Lin SS, O'Malley CD, Lui SW. Factors associated with missing birthplace information in a population-based cancer registry. *Ethn Dis* 2001; 11 (4): 598–605.

45. Haas A, Elliott MN, Dembosky JW, *et al.* Imputation of race/ethnicity to enable measurement of HEDIS performance by race/ethnicity. *Health Serv Res* 2019; 54 (1): 13–23.
46. Dembosky JW, Haviland AM, Haas A, *et al.* Indirect estimation of race ethnicity for survey respondents who do not report race/ethnicity. *Med Care* 2019; 57 (5): e28–33.
47. Gomez SL, Glaser SL. Quality of cancer registry birthplace data for Hispanics living in the United States. *Cancer Causes Control* 2005; 16 (6): 713–23.
48. Chen Y, Lin HY, Tseng TS, Wen H, DeVivo MJ. Racial differences in data quality and completeness: spinal cord injury model systems' experiences. *Top Spinal Cord Inj Rehabil* 2018; 24 (2): 110–20.
49. Sholle ET, Pinheiro LC, Adekkanattu P, *et al.* Underserved populations with missing race ethnicity data differ significantly from those with structured race/ethnicity documentation. *J Am Med Inform Assoc* 2019; 26 (8–9): 722–9.
50. Brown DP, Knapp C, Baker K, Kaufmann M. Using Bayesian imputation to assess racial and ethnic disparities in pediatric performance measures. *Health Serv Res* 2016; 51 (3): 1095–108.
51. Sherman RL. Address at diagnosis: place matters. *J Registry Manag* 2017; 44 (2): 76–7.
52. Cottrell EK, Hendricks M, Dambrun K, *et al.* Comparison of community-level and patient-level social risk data in a network of community health centers. *JAMA Netw Open* 2020; 3 (10): e2016852.
53. Lin G, Gray J, Qu M. Improving geocoding outcomes for the Nebraska Cancer Registry: learning from proven practices. *J Registry Manag* 2010; 37 (2): 49–56.
54. Dilekli N, Janitz A, Campbell J. Improved geocoding of cancer registry addresses in urban and rural Oklahoma. *J Registry Manag* 2020; 47 (1): 13–20.
55. Wilkins R. Use of postal codes and addresses in the analysis of health data. *Health Rep* 1993; 5 (2): 157–77.
56. Hughes AE, Pruitt SL. The utility of EMR address histories for assessing neighborhood exposures. *Ann Epidemiol* 2017; 27 (1): 20–6.
57. Ling C, Heck JE, Cockburn M, Liew Z, Marcotte E, Ritz B. Residential mobility in early childhood and the impact on misclassification in pesticide exposures. *Environ Res* 2019; 173: 212–20.
58. Armstrong BG. Effect of measurement error on epidemiological studies of environmental and occupational exposures. *Occup Environ Med* 1998; 55 (10): 651–6.
59. 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 (6 Suppl 1): S65–73.
60. McClure LA, Koru-Sengul T, Hernandez MN, *et al.* Availability and accuracy of occupation in cancer registry data among Florida firefighters. *PLoS One* 2019; 14 (4): e0215867.
61. Scott E, Hirabayashi L, Graham J, Krupa N, Jenkins P. Using hospitalization data for injury surveillance in agriculture, forestry and fishing: a crosswalk between ICD10CM external cause of injury coding and the Occupational Injury and Illness Classification System. *Inj Epidemiol* 2021; 8 (1): 6.
62. Silver SR, Tsai RJ, Morris CR, *et al.* Codability of industry and occupation information from cancer registry records: differences by patient demographics, casefinding source, payor, and cancer type. *Am J Ind Med* 2018; 61 (6): 524–32.
63. Armenti KR, Celaya MO, Cherala S, Riddle B, Schumacher PK, Rees JR. Improving the quality of industry and occupation data at a central cancer registry. *Am J Ind Med* 2010; 53 (10): 995–1001.
64. Polednak AP. Obtaining occupation as an indicator of patients' socioeconomic status in a population-based cancer registry. *J Registry Manag* 2005; 32 (4): 176–81.
65. Grundmeier RW, Song L, Ramos MJ, *et al.* Imputing missing race/ethnicity in pediatric electronic health records: reducing bias with use of U.S. Census location and surname data. *Health Serv Res* 2015; 50 (4): 946–60.
66. Wei II, Virnig BA, John DA, Morgan RO. Using a Spanish surname match to improve identification of Hispanic women in Medicare administrative data. *Health Serv Res* 2006; 41 (4 Pt 1): 1469–81.
67. Curriero FC, Kulldorff M, Boscoe FP, Klassen AC. Using imputation to provide location information for nongeocoded addresses. *PLoS One* 2010; 5 (2): e8998.
68. Kim J-S, Gao X, Rzhetsky A. RIDDLE: Race and ethnicity Imputation from Disease history with Deep LEarning. *PLoS Comput Biol* 2018; 14 (4): e1006106.
69. Henry KA, Boscoe FP. Estimating the accuracy of geographical imputation. *Int J Health Geogr* 2008; 7: 3.
70. Hibbert JD, Liese AD, Lawson A, *et al.* Evaluating geographic imputation approaches for zip code level data: an application to a study of pediatric diabetes. *Int J Health Geogr* 2009; 8 (1): 54.
71. Hsieh M-C, Pareti LA, Chen VW. Using NAPIIA to improve the accuracy of Asian race codes in registry data. *J Registry Manag* 2011; 38 (4): 190–5.
72. Hurley SE, Saunders TM, Nivas R, Hertz A, Reynolds P. Post office box addresses: a challenge for geographic information system-based studies. *Epidemiology* 2003; 14 (4): 386–91.
73. Kumar S, Liu M, Hwang S-A. A multifaceted comparison of ArcGIS and MapMarker for automated geocoding. *Geospat Health* 2012; 7 (1): 145–51.
74. Goldberg DW, Wilson JP, Knoblock CA, Ritz B, Cockburn MG. An effective and efficient approach for manually improving geocoded data. *Int J Health Geogr* 2008; 7 (1): 60.
75. Zhan FB, Brender JD, De Lima I, Suarez L, Langlois PH. Match rate and positional accuracy of two geocoding methods for epidemiologic research. *Ann Epidemiol* 2006; 16 (11): 842–9.
76. Weiss NS, Cooper SP, Socias C, Weiss RA, Chen VW. Coding of central cancer registry industry and occupation information: the Texas and Louisiana experiences. *J Registry Manag* 2015; 42 (3): 103–10.
77. Laberge M, Shachak A. Developing a tool to assess the quality of sociodemographic data in community health centres. *Appl Clin Inform* 2013; 4 (1): 1–11.
78. National Institute for Occupational Safety and Health (NIOSH). NIOSH Industry and Occupation Computerized Coding System (NIOCCS): Centers for Disease Control and Prevention (CDC). https://csams.cdc.gov/nioccs/ Accessed March 24, 2021.
79. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc* 2013; 20 (1): 144–51.
80. Polubriaginof FCG, Ryan P, Salmasian H, *et al.* Challenges with quality of race and ethnicity data in observational databases. *J Am Med Inform Assoc* 2019; 26 (8–9): 730–6.
81. Prosperi M, Min JS, Bian J, Modave F. Big data hurdles in precision medicine and precision public health. *BMC Med Inform Decis Mak* 2018; 18 (1): 139.
82. HL7. v3 Code System Race. Terminology Value Sets: HL7 International; 2018. https://www.hl7.org/fhir/v3/Race/cs.html Accessed March 24, 2020.
83. Pinto AD, Glattstein-Young G, Mohamed A, Bloch G, Leung F-H, Glazier RH. Building a foundation to reduce health inequities: routine collection of sociodemographic data in primary care. *J Am Board Fam Med* 2016; 29 (3): 348–55.
84. Vatsalan D, Christen P, Verykios VS. A taxonomy of privacy-preserving record linkage techniques. *Inf Syst* 2013; 38 (6): 946–69.
85. van der Lei J. Use and abuse of computer-stored medical records. *Methods Inf Med* 1991; 30 (2): 79e80.
86. Kressin NR, Chang B-H, Hendricks A, Kazis LE. Agreement between administrative data and patients' self-reports of race/ethnicity. *Am J Public Health* 2003; 93 (10): 1734–9.
87. Maizlish N, Herrera L. Race/ethnicity in medical charts and administrative databases of patients served by community health centers. *Ethn Dis* 2006; 16: 483–7.