





Using advanced racial and ethnic identity demographics to improve surveillance of work-related conditions in an occupational clinic setting

Andre G. Montoya-Barthelemy MD, MPH¹  | Karyn Leniek MD, MPH¹ | Emily Bannister MD, MPH¹ | Marcus Rushing MD, MS¹ | Fozia A. Abrar MD, MPH¹ | Tobias E. Baumann BA¹ | Madeleine Manly MA¹ | Jonathan Wilhelm BS¹ | Ashley Niece BA¹ | Scott Riester MD, PhD, MPH²  | Hyun Kim ScD³  | Jonathan Sellman MD, MPH¹ | Jay Desai PhD, MPH⁴ | Paul J. Anderson MD, MPH¹ | Ralph S. Bovard MD, MPH^{1,3} | Nicolas P. Pronk PhD⁵ | Zeke J. McKinney MD, MHI, MPH^{1,3,5} 

¹HealthPartners Department of Occupational and Environmental Medicine, St. Paul, Minnesota, USA

²Department of Occupational Medicine, Mayo Clinic, Rochester, Minnesota, USA

³School of Public Health, Division of Environmental Health Sciences, University of Minnesota, Minneapolis, Minnesota, USA

⁴Minnesota Department of Health, Health Promotion and Chronic Disease Division, St. Paul, Minnesota, USA

⁵HealthPartners Institute, Bloomington, Minnesota, USA

Correspondence

Andre Montoya-Barthelemy, MD MPH, HealthPartners Occupational and Environmental Medicine Residency, HealthPartners St. Paul Clinic, 205 S. Wabasha St., St. Paul, MN 55107, USA.
Email: agmb1214@gmail.com

Funding information

HealthPartners Institute, through the Regions Hospital Red Fund, St Paul, Minnesota

Abstract

Background: Although racial and ethnic identities are associated with a multitude of disparate medical outcomes, surveillance of these subpopulations in the occupational clinic setting could benefit enormously from a more detailed and nuanced recognition of racial and ethnic identity.

Methods: The research group designed a brief questionnaire to capture several dimensions of this identity and collected data from patients seen for work-related conditions in four occupational medicine clinics from May 2019 through March 2020. Responses were used to calculate the sensitivity and specificity of extant racial/ethnic identity data within our electronic health records system, and were compared to participants' self-reported industry and occupation, coded according to North American Industry Classification System and Standard Occupational Classification System listings.

Results: Our questionnaire permitted collection of data that defined our patients' specific racial/ethnic identity with far greater detail, identified patients with multiple ethnic identities, and elicited their preferred language. Response rate was excellent (94.2%, $n = 773$). Non-White participants frequently selected a racial/ethnic subcategory (78.1%–92.2%). Using our race/ethnicity data as a referent, the electronic health record (EHR) had a high specificity (>87.1%), widely variable sensitivity (11.8%–82.2%), and poorer response rates (75.1% for race, 82.5% for ethnicity, as compared to 93.8% with our questionnaire). Additional analyses revealed some industries and occupations disproportionately populated by patients of particular racial/ethnic identities.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *American Journal of Industrial Medicine* Published by Wiley Periodicals LLC

Conclusions: Our project demonstrates the usefulness of a questionnaire which more effectively identifies racial/ethnic subpopulations in an occupational medicine clinic, permitting far more detailed characterization of their occupations, industries, and diagnoses.

KEYWORDS

clinical surveillance, electronic health record, NAICS, occupational coding, occupational health, OIICS, racial and ethnic disparities, SOC system

1 | INTRODUCTION

Health inequality is a multifaceted, growing problem within the United States, encompassing diverse factors such as race/ethnicity, gender, socioeconomic status, and language.¹ Racial and ethnic occupational health disparities are obscured by an insufficient collection of demographic information. In particular, many surveillance systems rely on employer reporting, which is inconsistent between employers and undercounts both minor and chronic injuries.²⁻⁴ These issues are exacerbated among populations of vulnerable workers, such as undocumented immigrants and racial and ethnic minorities.^{5,6} The Occupational Safety and Health Administration (OSHA) does not mandate the collection of race/ethnicity data for injured workers.⁷ Similarly, race/ethnicity is an optional field in the Bureau of Labor Statistics Survey of Occupational Injury and Illnesses; these data were absent from 40% of days away from work cases in 2016.⁸ State-level programs, such as hospital discharge records, document measurement of race by observation, voluntary reporting, and inconsistent methodology.^{9,10}

Likewise, electronic health records (EHR) have limited fields for recording demographic data and these data are prone to inaccuracy. Klinger et al.¹¹ demonstrated that patients were 3% more likely to self-report Hispanic ethnicity and 6.4% more likely to report that they are Black than what is recorded in the EHR. Data for multiracial individuals were typically excluded from the EHR, with 30% of Whites, 37% of Hispanics, and 41% of African Americans self-reporting more than one race.¹¹ Similarly, a review of Veterans Affairs' EHR showed that race/ethnicity data was misclassified 15.7% of the time, which resulted either from data being missing entirely from the EHR or being incorrect.¹²

The importance of accurate racial/ethnic data was emphasized in 2019 by Riester et al.,¹³ who demonstrated disparate rates of occupational low back injuries in the Hispanic population in the Twin Cities. However, because of limited data on specific racial/ethnic compositions of the Hispanic population, no further inferences could be drawn as to exactly why these disparities exist. Similarly, the study by Riester et al. was limited in that industry, occupation, and detailed injury data regarding these injuries were not discretely captured, a common gap in clinical data collection.

Although the International Classification of Diseases, Tenth Revision (ICD-10) is used for clinical diagnostic coding consistently in the United States and has codes to describe injury details, these data

are frequently imprecise or missing data beyond a primary diagnosis.¹⁴

Precise and standardized ontologies for industry and occupation exist, such as the North American Industrial Classification System (NAICS) and Standard Occupational Classification (SOC), but these ontologies have not been commonly implemented in clinical settings using an EHR. The importance of documenting work and its components as part of evaluating social determinants of health within an EHR for individuals and populations was discussed at length by Schmitz and Forst.¹⁵ Their work evaluates a computerized autocoding system developed by the National Institute for Occupational Safety and Health (NIOSH) to generate a code representing industries and occupations, the National Industry and Occupation Computerized Coding System (NIOCCS). The tools described above, used in combination, represent the dimensions of work-related health that are needed in clinical data collection to truly implement occupational population health surveillance.

It is the responsibility of healthcare providers to understand the compositions and needs of the populations they serve, which is not possible without appropriate data quality. This is vital for providers in diverse areas, such as the Twin Cities metropolitan area. Though overall smaller than the national average, (8% vs. 14%), Minnesota's population of foreign-born individuals is outpacing national growth, doubling nationally but tripling in Minnesota.¹⁶ Furthermore, the Minneapolis-St. Paul area is home to both the largest Hmong and Somali populations in the nation.^{17,18} Taken together, these gaps in data quality and evidence of existing occupational health disparities indicate a unique opportunity for twin cities-based occupational medicine practitioners to better understand these disparities in regard to work-related conditions and intervene to mitigate them. As such, this study aimed to improve race/ethnicity data quality relative to existing EHR data, as a step toward identifying and understanding differential injury risks based in populations with various demographic and workplace factors.

2 | MATERIALS AND METHODS

2.1 | Study design, setting, and population

To better inform shared clinical decision-making, this cross-sectional observational study is designed to improve and evaluate the quality

of data available to occupational medicine providers and their patients by collecting, coding, and linking data from multiple data sources. The study settings include four dedicated occupational medicine clinics within HealthPartners, a large integrated health system in Minnesota. These clinics exclusively provide care for work-related injuries and illnesses originating both from independent community referral and from contracted employers, pre-placement and surveillance examinations, and consultations related to environmental illness, work ability, and determination of work-related disability. Study participants were 18 years and older and presented to one of these clinics for new workers' compensation injuries from May 1, 2019 through March 31, 2020. Data collection was discontinued on March 31, 2020 with the advent of the COVID-19 global pandemic and unanticipated changes to clinic processes. This study protocol was exempt from full review by the HealthPartners Institutional Review Board and did not meet the definition of human subjects research. It was instead determined to be a quality improvement project, collecting data that was already solicited through existing processes, and posing only minimal additional burden to participants.

2.2 | Data collection

This study merges four primary data sets: (1) detailed self-reported demographic data; (2) existing EHR data (Epic Systems Corporation); (3) coded industry (via North American Industry Classification System [NAICS]) and occupation (via Standard Occupational Classification [SOC] system); (4) coded data on the specific patient injury (via Occupational Injury & Illness Classification System [OIICS]). The merged data establishes the conceptual foundation of an aggregated data system to enhance occupational injury surveillance by allowing for the identification of inequities in injuries amongst subpopulations not otherwise identifiable by discrete data collected through existing EHR data structures and clinical data collection practices.

A new paper-based demographic form was developed to capture self-identified data in a clinical occupational medicine setting (Figure 1). As part of development, the form was reviewed by Hmong, Spanish, and Somali interpreters and patients, piloted, and revised to assure accuracy of answers. Marital status, education level, race/ethnicity, country of birth, and language(s) spoken at home were included on the new demographic form. The paper form was a one single-sided page to reduce the likelihood of missed responses and minimize the time required to complete the form.

The new demographic form contained the same selections for marital status as are available in the EHR. The questionnaire sought to engage the many facets of race, ethnicity, and cultural identity by gathering multiple related data points in a more accessible, open, and relevant way than is permitted by the current EHR system. For example, the form asks broadly "which of the following do you consider yourself?" (rather than simply asking for "Race/Ethnicity"), for "country of birth" (rather than "country origin" as in the EHR), and for language as "what language(s) do

you speak at home?" Also, unlike in the EHR, patients could select as many race/ethnicity and language categories as they preferred. Finally, the questionnaire was completed by the participants themselves, rather than being solicited by clinic staff at the time of registration.

Multiple options were provided for both race and ethnicity. For instance, Hispanic ethnicity and race categories were combined into one question to eliminate the common mistake of patients checking "Yes" to Hispanic and then specifying "Hmong" in the comments field. The final race categories were White, American Indian/Alaskan Native, Asian, Black/African/African American, Native Hawaiian/Other Pacific Islander, and "Other background not listed above," with allowance for free-text entry of race/ethnicity. Subcategories chosen for Asian, Black, and Hispanic identities based on 2010 U.S. Census data for the Minneapolis–St. Paul metropolitan area, with an additional "Other" free-text field after each category listing, similar to the paper form used for the United Kingdom census.¹⁹

The 10 most commonly spoken languages and 10 most common countries of birth for foreign-born individuals in the Minneapolis–St. Paul metropolitan area were listed, followed by an option to choose "Other" and enter free-text. Languages and countries were limited to the 10 most common to minimize the length of the form, with an option for "Other" and space for free-text entry.

The new demographic form was administered to all patients attending an initial workers' compensation injury visit. Patients were provided an informational handout detailing the project and signed an opt-out informed consent. If an interpreter was present for the visit, the interpreter read aloud the questions and answer choices in the patient's primary language. Demographic form data were manually entered into the REDCap database (Research Electronic Data Capture), with patient identifiers of the patient medical record number (MRN) and the date of their clinical encounter.^{20,21} REDCap is a secure, web-based software platform designed to support data capture for research studies, providing (1) an intuitive interface for validated data capture; (2) audit trails for tracking data manipulation and export procedures; (3) automated export procedures for seamless data downloads to common statistical packages; (4) procedures for data integration and interoperability with external sources. Participants who declined to fill out forms were included in the database with only that routine data which had already been collected in the EHR, but no further data were solicited or collected. Overall, they constituted a very small proportion of the total.

2.3 | EHR data

Routine data collection in the HealthPartners EHR includes age, sex, weight, height, in addition to individual demographics of race, ethnicity, marital status, education level, country of origin, and preferred language, as well as occupational demographic data including employer name and patient occupation. Demographic and occupational data within the EHR are accessible within a "Demographics" activity of the EHR and can be entered by clinical and nonclinical staff as part



Filling out this form is optional.

1. What is your marital status?
 - Divorced
 - Legally separated
 - Married
 - Significant other/partner
 - Single/never married
 - Widowed

2. What is your highest level of education?
 - Not a high school graduate
 - High school diploma or GED
 - Technical certificate or diploma
 - Associate degree
 - Bachelor degree
 - Advanced degree

3. Which of the following do you consider yourself?
(Mark all that apply)
 - White
 - American Indian or Alaskan Native
 - Asian
 - Chinese
 - Hmong
 - Indian
 - Laotian
 - Thai
 - Vietnamese
 - Other Asian background: _____
 - Black, African, or African American
 - African American
 - Black
 - Ethiopian
 - Liberian
 - Nigerian
 - Somali
 - Other Black/African background: _____
 - Hispanic, Latino, or Latina
 - Mexican
 - Puerto Rican
 - Other Hispanic/Latino background: _____
 - Native Hawaiian or other Pacific Islander
 - Other background not listed above: _____

Please return form to the front desk staff

4. Where were you born?
 - United States
 - Ethiopia
 - India
 - Laos
 - Liberia
 - Mexico
 - Somalia
 - Thailand
 - Vietnam
 - Other country: _____

5. What language(s) do you speak at home?
(Mark all that apply)
 - English
 - Amharic
 - Hmong
 - Oromo
 - Somali
 - Spanish
 - Vietnamese
 - Other language: _____

FIGURE 1 Newly created single-page demographic form

of or independent of standard clinical workflows, with the exception of educational level which is only accessible elsewhere in the EHR.

EHR data for clinical encounters in the study clinics were extracted periodically from the EHR by the HealthPartners Institute Research Informatics and Information Services team by matching MRN and encounter date from collected demographic forms and were imported into the study REDCap database. Diagnosis data, coded as one or more ICD-10 codes, was also extracted with encounter demographic data, and was used to validate injury coding in cases where needed, but was not otherwise included in the analysis.

2.4 | Industry, occupation, and injury data coding

Injury, industry, occupation data were manually coded by one of the multiple research assistants who had undergone specialized training with NIOSH. These were combined within the REDCap database with clinical diagnoses and free-text employer name and job title, each extracted from the clinical documentation. Industry, occupation, and injury data were left blank if they could not be coded based on available data, and the compiled results reviewed by one of the study investigators.

Injury data were coded with one code each per work injury encounter according to OIICS, a hierarchical coding system developed by the United States Bureau of Labor Statistics in 1998. OIICS codes for type (nature) of injury, part of body, source of injury, and mechanism of injury, indicating greater degrees of specificity with increased length of the code. OIICS data were manually assigned by chart review, then were re-reviewed after initial collection to ensure data quality. Industry and occupation were coded using NIOCCS, a computerized autocoding system, to generate NAICS codes. Initial attempts at autocoding with NIOCCS had only limited success as employer names did not contain enough data for sufficient coding, resulting in primarily manual coding of industry and occupation. NAICS data describing industry were reported at the level of industry sector (one-digit codes) and subsector (two-digit codes).

2.5 | Data analysis

Data were extracted from REDCap into Microsoft Excel 2013 (Microsoft Corporation), and reported as frequencies and percentages. Primary descriptive analysis included evaluation of demographic form data collection and reporting, and comparison of demographic form data quality with data collected in the EHR. Data were considered “unknown” in the EHR if the field was marked as “Unknown” or if the field was left blank in the EHR or on the form (including on “declined” forms). Free-text data that had an available discrete selection category were recoded into the appropriate category.

Sensitivity and specificity were calculated for EHR demographic data relative to the patient-reported demographic form data (i.e., gold standard). True-positives (TP) and -negatives (TN) for each demographic element were cases where the EHR and new form were in

agreement about the specific characteristic. “False-positive” describes instances where the EHR was “positive” and the self-report form was “negative” (patient denied this characteristic) while “false-negative” describes instances where the EHR was “negative” and the self-report form was “positive” (patient endorsed a characteristic not listed in the EHR). Reliability of EHR data relative to demographic form data was calculated with Cohen's κ , where the observed agreement was calculated as TP + TN, while the expected agreement was calculated using marginal sums.²² Descriptive analyses included race/ethnicity data for the categories of Asian, Black/African American, and Hispanic, distributions of detailed languages and birth countries, and industry/occupation data as coded according to the NAICS/SOC framework.

3 | RESULTS

During the study period, the four occupational medicine clinics documented 8151 encounters, of which 3087 (37.9%) were completed for workers' compensation care, and of which 1118 (13.7%) were new patient evaluations. Collection of the new demographic forms was attempted during 821 of those new evaluations, comprising 73.4% (821/1118) of new patient work injury evaluations. Of these 821 forms, 773, were marked with data of any kind and were included in the analysis, for an overall response rate of 94.2%. A staged rollout of the questionnaire across four separate clinics, especially at the beginning of data collection, likely explains the incomplete collection across all new evaluations.

Of those forms that were completed, individual data elements (marital, education, race/ethnicity, country, language) were well represented, with greater than 98% of each element completed (see Table 1).

The patient population ranged from 20 to 83 years of age (mean: 44.3 ± 12.6 years), of whom 53.0% ($n = 435$) were men.

Of individuals who reported race/ethnicity (759/773 total completed forms), the most common were White ($n = 510$, 67.2%), Black/African/African American ($n = 146$, 19.2%), Hispanic/Latino/Latina ($n = 73$, 9.6%), and Asian ($n = 51$, 6.7%). Overall, the agreement between the questionnaire and EHR data regarding major race/ethnicity groups was widely variable, highest for Black/African American ($\kappa = 0.80$) and lowest for American Indian/Alaska Native ($\kappa = 0.20$).

Respondents selected a subcategory within the Asian race/ethnicity in 47 of 51 cases (92.2%), within the Black/African/African American category in 114 of 146 cases (78.1%), and within the Hispanic/Latino/Latina category in 67 of 73 cases (91.8%), with the largest number of respondents identifying as Indian, African American, and Mexican, respectively. One Asian respondent selected two races/ethnicities, five Black/African American respondents selected two, one Black/African American respondent selected three, and one Hispanic respondent selected two races/ethnicities. They also made frequent use of the “Other” write-in option, with 21 Black (14.4%), 27 Hispanic (37.0%), and 18 Asian (35.2%) respondents selecting an “Other” subgroup, yielding 28 additional ethnicities not specifically

TABLE 1 Frequency of collected demographic form elements that were completed

Form Element	Completed (n)	Percent total (n = 821, %)	Percent completed (n = 773, %)
Marital data	765	93.2	99.0
Education data	761	92.7	98.5
Race data	759	92.5	98.2
American Indian or Alaskan Native	17	2.1	2.2
Asian	51	6.2	6.6
Black, African, or African American	146	17.8	18.9
Hispanic, Latino, Latina	73	8.9	9.4
Native Hawaiian	5	0.6	0.7
White	510	62.1	66.0
Other selected	16	2.0	2.1
Other filled	14	1.7	1.8
Multiple races selected (excluding Other)	44	5.4	5.7
Multiple races selected (including Other)	53	6.5	6.9
Country data	760	92.6	98.3
Other selected	67	8.2	8.7
Other filled	63	7.7	8.2
Language data	760	92.6	98.3
Other selected	45	5.5	5.8
Other filled	43	5.2	5.6

Note: "Other selected" indicates those who marked the discrete selection of "Other" while "Other filled" indicates those who filled out free-text entries within form elements.

listed on the form. Details of self-reported racial/ethnic identification and subgroups are detailed in Tables 2 and 3.

The full list of languages entered by respondents may be found in Supporting Information Appendix Table SI, and the full list of countries from the EHR and on the demographic may be found in Supporting Information Appendix Table SII. Within the study population, 90 individuals (11.6% of completed forms) selected multiple languages, of which 30 included the selection of "Other." Education data, which was only available in the paper-based demographic form, is depicted in Figure 2.

On the basis of EHR race data, the 48 respondents who declined to complete the demographic form were documented as 45.8% White, 18.8% Black/African American, 8.3% Asian, 2.1% American Indian/Alaska Native, and 25.0% with no race data entered in the EHR. With respect to ethnicity as defined in the EHR, 72.9% were non-Hispanic and 2.1% were Hispanic ($n = 1$), while 25.0% had no ethnicity specified (Table 3).

For the 821 patients from whom these forms were collected, marital data was available in the EHR in 765 cases (93.2%), race data in 617 cases (75.2%), ethnicity data in 677 cases (82.5%), and language data in 803 cases (97.8%). Education data was not retrievable from the EHR in any case.

The collected demographic form data were compared to collected EHR data (Table 2). The EHR demonstrated high specificity for marital, race, ethnicity, and language data elements, with lower specificity for single marital status (73.4%), unknown race (75.2%), unknown ethnicity (82.9%), and English language (37.6%). Sensitivity for EHR data ranged widely for all data elements. The EHR lacked data for marital status, race, ethnicity, and language in 27.7%, 24.9%, 17.5%, and 2.2% of cases, versus 7.2% for marital status, as compared to 6.2% for each race, ethnicity, and language from the demographic form (Table 2).

Country data in the EHR was not available for $n = 329$ patients (321 blank field, 6 "Other," and 2 "Patient refused to answer"). Country data within the EHR as compared to the demographic form demonstrated extremely high specificity (>90%) for all countries, but wide-ranging sensitivity (Supporting Information Appendix Table SII).

Industry and occupation were coded for all patients when possible; industry was unable to be coded in eight cases and occupation in 74 cases, yielding 813 with industry coding and 747 with occupational coding, shown in Table 4. The top four industries represented 485 of 813 (59.7%) of patients, comprising health care and social assistance (NAICS sector 62; $n = 149$, 18.3%), manufacturing

TABLE 2 Electronic health record demographic data in comparison to collected demographic form data

EHR marital status	EHR (n = 821) ^a	Form (n = 765) ^a	TP (% EHR) ^b	TN (% EHR) ^b	FP (% EHR) ^b	FN (% EHR) ^b	Sensitivity (%)	Specificity (%)	κ
Divorced	39	89	3.7	88.1	1.1	7.2	33.7	98.8	0.43
Legally separated	<5	11	<0.6	98.2	<0.6	1.3	0.0	99.5	0.01
Married	223	314	22.0	56.6	5.1	16.2	57.6	91.7	0.52
Other ^c	<5								
Significant other/partner	20	97	1.5	87.2	1.0	10.4	12.4	98.9	0.17
Single	300	245	17.9	51.5	18.6	11.9	60.0	73.4	0.31
Unknown ^d	227	104	2.1	68.6	3.8	25.6	7.5	94.8	0.03
Widowed	5	9	0.6	98.9	<0.6	<0.6	55.6	100	0.71
EHR race	EHR (n = 821)	Form (n = 759)	TP	TN	FP	FN	Sensitivity	Specificity	
American Indian or Alaska Native	<5	17	<0.6	97.8	<0.6	1.8	11.8	99.9	0.20
Asian	31	51	3.3	93.3	<0.6	2.9	52.9	99.5	0.64
Black or African American	127	146	13.9	80.6	1.6	3.9	78.1	98.1	0.80
White	440	510	48.7	33.0	4.9	13.4	78.4	87.1	0.63
More than 1 race ^e	8	53	<0.6	93.1	<0.6	6.0	7.6	99.5	0.12
More than 1 race ^f	8	33	<0.6	95.5	<0.6	3.5	12.1	99.5	0.18
Some other race	8	16	<0.6	97.2	0.9	1.8	6.3	99.1	0.07
Unknown ^d	204	48	1.5	70.8	23.4	4.4	25.0	75.1	0.00
EHR ethnicity	EHR (n = 821)	Form (n = 759)	TP	TN	FP	FN	Sensitivity	Specificity	
Hispanic	54	73	5.8	90.4	0.7	3.0	65.8	99.2	0.74
Non-Hispanic	623	748	74.9	7.9	1.0	16.2	82.2	89	0.40
Unknown ^d	144	48	1.5	78.1	16.1	4.4	25.0	82.9	0.04
EHR language	EHR (n = 821)	Form (n = 760)	TP	TN	FP	FN	Sensitivity	Specificity	
English	749	696	81.7	5.7	9.5	3.0	96.4	37.6	0.41
Amharic	<5	11	<0.6	98.7	<0.6	1.1	18.2	100	0.31
Hmong	<5	17	<0.6	97.7	<0.6	1.8	11.8	99.8	0.18
Oromo	6	10	<0.6	98.5	<0.6	0.7	40.0	99.8	0.50
Somali	6	14	<0.6	98.2	<0.6	1.1	35.7	99.9	0.50
Spanish	22	60	2.4	92.4	<0.6	4.9	33.3	99.7	0.47
Vietnamese	<5	<5	<0.6	99.3	<0.6	<0.6	50.0	99.8	0.50
Other	<5	45	<0.6	94.5	<0.6	5.4	2.2	100	0.04
American Sign Language	<5	5	<0.6	99.4	<0.6	<0.6	40.0	100	0.50
Arabic	<5	<5	<0.6	99.6	<0.6	<0.6	33.3	100	0.00
Cambodian	<5	<5	<0.6	99.8	<0.6	<0.6	0.0	99.9	0.33
French	<5	5	<0.6	99.4	<0.6	<0.6	20.0	100	0.57
Swahili	<5	<5	<0.6	99.6	<0.6	<0.6	33.3	100	0.50
Tibetan	<5	<5	<0.6	99.9	<0.6	<0.6	0.0	99.9	0

(Continues)

TABLE 2 (Continued)

EHR language	EHR (n = 821)	Form (n = 760)	TP	TN	FP	FN	Sensitivity	Specificity	
Tigrinya	<5	<5	<0.6	99.8	<0.6	<0.6	100.0	99.9	0.666
Unknown ^d	18	61	<0.6	90.7	1.8	7.1	4.9	98.1	0.04

Abbreviations: EHR, elections within the electronic health record, TP + FP; Form, elections on the collected demographic form, TP + FP, total of form column may be greater than n = 821 due to ability to choose multiple selections on form; FN, false-negative, form positive, EHR negative; FP, false-positive, EHR positive, form negative; TN, true-negative, agreement between form and EHR; TP, true-positive, agreement between form and EHR.

^aCell counts of less than five were reported as such within the table to minimize identifiability of smaller subpopulations within the study; all such results are ordered according to the actual frequencies, with equal frequencies ordered in alphabetical order.

^bFor cell counts of less than five, percentages for TP/TN/FP/FN were reported as <0.6% (5/821).

^c"Other" not a selection on demographic form for marital status.

^d"Unknown" or blank in EHR field, form declined, or form field blank.

^eForm comparison if selected more than one race on the form.

^fForm comparison if selected more than one race on form, only including races from EHR.

TABLE 3 Demographic form subcategories selected for race/ethnicity categories of Asian, Black/African American, and Hispanic

Asian (n = 51)	Specific race selected (n = 47) ^{a,b}		Black/African American (n = 146)	Specific race selected (n = 114) ^{a,b}		Hispanic (n = 73)	Specific race selected (n = 67) ^{a,b}	
	n	%		n	%		n	%
Hmong	18	35.3	African American	43	29.5	Mexican	35	48.0
Vietnamese	5	9.8	Black	30	20.6	Other ^c	7	9.6
Filipino	5	9.8	Ethiopian	13	8.9	Puerto Rican	<5	<9.8
Chinese	<5	<9.8	Somali	12	8.2	Hispanic	<5	<9.8
Korean	<5	<9.8	Liberian	10	6.9	Brazilian	<5	<9.8
Other ^c	<5	<9.8	Nigerian	<5	<9.8	Colombian	<5	<9.8
Indian	<5	<9.8	African	<5	<9.8	Cuban	<5	<9.8
Laotian	<5	<9.8	Eritrean	<5	<9.8	Dominican	<5	<9.8
Cambodian	<5	<9.8	Moorish	<5	<9.8	Ecuadorian	<5	<9.8
Japanese	<5	<9.8	Other ^c	<5	<9.8	Guatemalan	<5	<9.8
Eritrean	<5	<9.8	Black American	<5	<9.8	Mexican American	<5	<9.8
Korean American	<5	<9.8	East African	<5	<9.8	Costa Rican	<5	<9.8
South Korean	<5	<9.8	Guyanese	<5	<9.8	Honduran	<5	<9.8
Tibetan	<5	<9.8	Kenyan	<5	<9.8	Latino American	<5	<9.8
			Nepali	<5	<9.8	Spanish	<5	<9.8

^aTotals of columns may be greater than the number of individuals selecting specific ethnicities as individuals may have indicated more than one specific ethnicity.

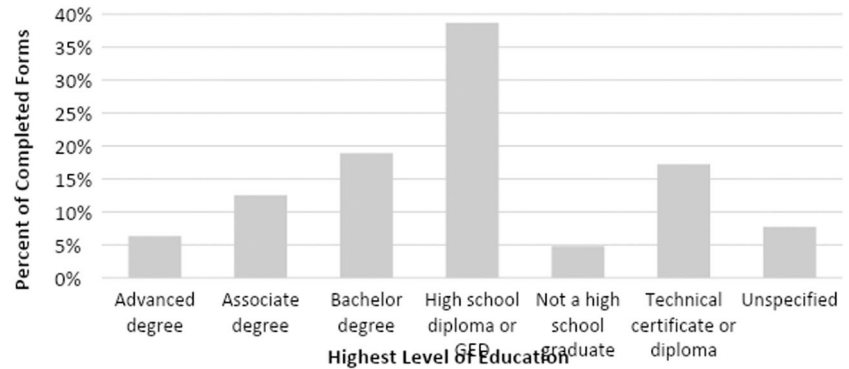
^bCell counts of less than five were reported as such within the table to minimize likelihood of identifiability of smaller subpopulations within the study; all such results are ordered according to the actual frequencies, with equal frequencies ordered in alphabetical order.

^cOther selected without any further clarification, such as a free-text entry, on the form.

(31–33; n = 145, 17.8%), educational services (61; n = 124, 15.2%), and public administration (92; n = 67, 8.2%). The top four occupations accounted for 354 of 747 patients (43.1%), and are production (SOC major group 51; n = 149, 19.9%), transportation and material moving (53; n = 91, 12.2%), healthcare practitioners and technical (29; n = 57; 7.6%), and healthcare support (31; n = 57, 7.6%).

Finally, OIICS data for nature of injury and part of body were coded for 762 patients (92.8% of all patients), for injury source in 693 patients (84.4%), for secondary source in 109 patients (13.3%), and for mechanism of injury in 759 patients (92.5%). The most common nature of injury was "traumatic injuries and disorders" (OIICS nature division 1; n = 648, 85%), of which 35.3% (OIICS nature major group

FIGURE 2 Highest level of education reported on the demographic form ($n = 761$)



12, $n = 229$) were “traumatic injuries to muscles, tendons, ligaments, joints, and so on.” The most common part of the body injured was the upper extremity (OIICS part division 4; $n = 317$, 41.6%), the most common injury source was “persons, plants, animals, and minerals” (OIICS source division 5; $n = 268$, 38.7%), and the most common mechanism of injury was “overexertion and bodily reaction” (OIICS event/exposure division 7; $n = 241$, 31.8%).

4 | DISCUSSION

This study was successful in collecting detailed demographic data on new occupational injury patients as evidenced by the high response rates of the demographic forms and of individual questions, suggesting the length of the form and the content of the form were not a barrier to completion. The primary success of this data collection effort was the ability to collect detailed race and ethnicity data in this occupational medicine population, not achieved with EHR data collection practices to date.

Another way in which this independent demographic data collection effort was successful was in identifying gaps in existing EHR data collection practices as they pertain to the collection of demographic data (i.e., marital status, education, race, country, language). For example, no educational level data were captured in the EHR for the patient population in this study because the field for entry of educational attainment cannot be entered during registration in its present implementation. For the other demographic data (aside from education), data collection from the EHR demonstrated generally very high rates of specificity as compared to demographic form data, whereas sensitivity was mixed across all data collection. This variable sensitivity reflects higher false-negative rates for EHR data collection and suggests that individuals may have been limited in only being able to make a single selection in the EHR or that clinical processes may not be accurately capturing some demographic characteristics. Below, we describe specific analyses related to the individual demographic data points.

The form design largely succeeded in eliciting race/ethnicity data as patients commonly selected subcategories within each race, as shown by the 270 Black, Hispanic, and Asian patients (the three races with listed subcategories) having selected an ethnicity

subcategory in 228 cases (84.4%). The fact that 28 ethnicities were written in free-text on the form demonstrates that the diversity of patients served by this clinic is greater than could neatly fit on a one-page form. These data suggest that the method of maximizing the available options while also keeping selections simple and concise was successful in providing detailed and useful data.

An excellent demonstration of value to this added detail is exemplified by Hmong ethnicity, which EHR data is unable to capture effectively. Hmong identity is unique in that it is not strongly tied to one country of origin; Hmong peoples are dispersed across China, Vietnam, Thailand, and Laos, as well as diaspora communities in the United States.²³ The only field that would capture Hmong identity in the EHR is the language field, yet in this it failed almost every time, registering 15 false-negatives and only two true-positives for a specificity of 11.8% ($\kappa = 0.18$; see Table 2).

Self-identification of race and ethnicity, particularly in the African American/Black community, yielded interesting results. Both African American and Black were included as ethnic options within the main category, and respondents chose each of these options at comparable rates (29.5% African American, 20.6% Black, and only 2.7% both). For individuals that selected African American/Black and specified a country-specific ethnicity (e.g., Nigerian), there was enormous variability of selecting African American, Black, neither, or both in addition (data not shown). Although they are literally “Black” and “African American,” African immigrants may largely perceive themselves as belonging to a separate and distinct cultural group. Benjamin Okonofua, Professor of Sociology at University of Benin, demonstrates that African immigrants often define their identity based on distinct languages, sociocultural heritage, and national origin, and prefer to avoid racial categorization as Black or African American, groups with a specific history of discrimination and oppression in the United States.²⁴

These data strongly support the value of a data collection method, which better defines the nuances of racial and ethnic identity. It would also better indicate community groups best positioned for collaboration. While a full sociological analysis is outside the scope of this article, similar benefits could be gained by better understanding within-groups differences among respondents classified by EHR as “Hispanic/Latino” or “Asian.”

TABLE 4 Industries and occupations represented, in total and by major race/ethnicity categories (in the order listed on the demographic form)

NAICS sector (NAICS code)	Total ^a n	White ^a		Native American ^a		Asian ^a		Black ^a		Hispanic ^a	
		n	%	n	%	n	%	n	%	n	%
Agriculture, Forestry, Fishing, and Hunting (11)	6	5	83.3	<5	NR	<5	NR	<5	NR	<5	NR
Mining, Quarrying, and Oil and Gas Extraction (21)	<5	<5	NR	<5	NR	<5	NR	<5	NR	<5	NR
Utilities (22)	<5	<5	NR	<5	NR	<5	NR	<5	NR	<5	NR
Construction (23)	30	20	66.7	<5	NR	<5	NR	6	20.0	<5	NR
Manufacturing (31–33)	145	91	62.8	<5	NR	13	9.0	16	11.0	14	9.7
Wholesale Trade (42)	37	25	67.6	<5	NR	<5	NR	<5	NR	<5	NR
Retail Trade (44–45)	50	32	64.0	<5	NR	<5	NR	15	30.0	<5	NR
Transportation and Warehousing (48–49)	53	28	52.8	<5	NR	<5	NR	10	18.9	<5	NR
Information (51)	16	13	81.3	<5	NR	<5	NR	<5	NR	<5	NR
Finance and Insurance (52)	4	<5	NR	<5	NR	<5	NR	<5	NR	<5	NR
Real Estate and Rental and Leasing (53)	4	<5	NR	<5	NR	<5	NR	<5	NR	<5	NR
Professional, Scientific, and Technical Services (54)	16	12	75.0	<5	NR	<5	NR	<5	NR	<5	NR
Management of Companies and Enterprises (55)	<5	<5	NR	<5	NR	<5	NR	<5	NR	<5	NR
Administrative and Support and Waste Management and Remediation Services (56)	42	24	57.1	<5	NR	<5	NR	7	16.7	5	11.9
Educational Services (61)	124	88	71.0	<5	NR	6	4.8	21	16.9	6	4.8
Healthcare and Social Assistance (62)	149	70	47.0	<5	NR	13	8.7	47	31.5	11	7.4
Arts, Entertainment, and Recreation (71)	18	16	88.9	<5	NR	<5	NR	<5	NR	0	0.0
Accommodation and Food Services (72)	38	14	36.8	<5	NR	<5	NR	10	26.3	13	34.2
Other services (except Public Administration) (81)	9	6	66.7	<5	NR	<5	NR	<5	NR	<5	NR
Public Administration (92)	67	52	77.6	<5	NR	<5	NR	<5	NR	6	9.0
Unknown ^b	8	6	75.0	<5	NR	<5	NR	<5	NR	<5	NR
SOC occupation (SOC code)											
Management (11)	<5	<5	NR	<5	NR	<5	NR	<5	NR	<5	NR
Business and Financial Operations (13)	6	<5	NR	<5	NR	<5	NR	<5	NR	<5	NR
Computer and Mathematical (15)	6	<5	NR	<5	NR	<5	NR	<5	NR	<5	NR
Architecture and Engineering (17)	17	14	82.4	<5	NR	<5	NR	<5	NR	<5	NR
Life, Physical, and Social Science (19)	15	14	93.3	<5	NR	<5	NR	<5	NR	<5	NR
Community and Social Service (21)	13	11	84.6	<5	NR	<5	NR	<5	NR	<5	NR
Legal (23)	<5	<5	NR	<5	NR	<5	NR	<5	NR	<5	NR
Educational Instruction and Library (25)	53	36	67.9	<5	NR	<5	NR	10	18.9	<5	NR
Arts, Design, Entertainment, Sports, and Media (27)	<5	<5	NR	<5	NR	<5	NR	<5	NR	<5	NR
Healthcare Practitioners and Technical (29)	57	37	64.9	<5	NR	7	12.3	8	14.0	<5	NR
Healthcare Support (31)	57	18	31.6	<5	NR	<5	NR	25	43.9	<5	NR
Protective Service (33)	42	33	78.6	<5	NR	<5	NR	<5	NR	6	14.3
Food Preparation and Serving Related (35)	34	11	32.4	<5	NR	<5	NR	13	38.2	10	29.4
Building and Grounds Cleaning and Maintenance (37)	45	20	44.4	<5	NR	<5	NR	10	22.2	7	15.6
Personal Care and Service (39)	23	17	73.9	<5	NR	<5	NR	<5	NR	<5	NR
Sales and Related (41)	16	12	75.0	<5	NR	<5	NR	<5	NR	<5	NR

TABLE 4 (Continued)

NAICS sector (NAICS code)	Total ^a	White ^a		Native American ^a		Asian ^a		Black ^a		Hispanic ^a	
	n	n	%	n	%	n	%	n	%	n	%
Office and Administrative Support (43)	44	24	54.6	<5	NR	5	11.4	8	18.2	<5	NR
Farming, Fishing, and Forestry (45)	11	11	100.0	<5	NR	<5	NR	<5	NR	<5	NR
Construction and Extraction (47)	35	21	60.0	<5	NR	<5	NR	<5	NR	<5	NR
Installation, Maintenance, and Repair (49)	23	19	82.6	<5	NR	<5	NR	<5	NR	<5	NR
Production (51)	149	93	62.4	<5	NR	11	7.4	20	13.4	15	10.1
Transportation and Material Moving (53)	91	49	53.9	<5	NR	7	7.7	22	24.2	<5	NR
Military Specific (55)	<5	<5	NR	<5	NR	<5	NR	<5	NR	<5	NR
Unknown ^b	74	53	71.6	<5	NR	6	8.1	0	0.0	0	0.0

Abbreviations: NAICS, North American Industrial Classification System; SOC, Standard Occupational Classification.

^aCell counts of less than five (other than zero) were reported as such within the table to minimize the likelihood of identifiability of smaller subpopulations within the study; all such results are ordered according to the actual frequencies, with equal frequencies ordered in alphabetical order. Corresponding percentages listed as NR (not reportable) as frequencies identifiable from percentages.

^bUnknown: unable to be coded based on available industry/occupation data within char.

Language data demonstrated an interesting discrepancy between the form and the EHR data in terms of English language as opposed to other languages. Although language data were more often available in the EHR relative to the demographic form, a greater proportion of individuals in the EHR were assigned English as their preferred language than on the demographic form, which was the only element in Table 2 for which EHR sensitivity outranked specificity. Because the demographic form allowed for multiple selections of language, it is possible the demographic form more easily permitted selection of non-English languages. However, assignment of the English language on the part of clinical staff if a patient demonstrates competency (in initial interactions in the clinic) could inflate those numbers as well. In addition, the question on the demographic form asks “what languages do you speak at home?” perhaps leading each set of data to reflect the context in which the data was collected. In other words, a patient would more likely accommodate the English-speaking clinic but would prefer to speak a family language at home. English language data yielded a low specificity (37.6%) and high sensitivity (96.4%).

Country of origin was also much more consistently collected with the new demographic form (92.5%, $n = 759$) than in the dedicated field in the EHR ($n = 492$, 59.9%). Reliability of country data was mixed overall but was notably low ($\kappa = 0.34$) for the United States.

Marital status in the EHR was highly specific with mixed sensitivity, similar to most other collected fields; it was observed that there were a high proportion of false-negatives for “Significant Other/Partner;” but this effect may be related to the fact that marital/relationship status may be transient, may not be re-evaluated at subsequent clinical visits, or queried at the time of collection between “married” versus “single” options. Reliability of marital status as measured by κ was at best moderate (0.71) in the case of “Widowed,” but otherwise was weak (<0.60) or worse.

Table 4 breaks down the sample by coded occupation, using both NAICS and SOC codes, demonstrating that some ethnic groups may have disproportionate representation in certain labor-intense

occupations. For example, the SOC coding system data demonstrate that White people are overrepresented within “Healthcare Practitioners and Technical,” representing higher-paid, more prestigious, and more highly trained positions (e.g., physicians, respiratory therapists, or radiation technicians), and very poorly represented within “Healthcare Support,” which are generally more poorly compensated and have more labor-intensive jobs (e.g., personal care assistants and home health aides). Barriers to education, credentials, and networking faced by minority communities, especially Black and Hispanic populations, likely drive them disproportionately into Healthcare Support jobs with higher rates of injuries, especially strains and chronic pain caused by repetitive strain or overexertion.^{25,26} Black and Hispanic people are similarly overrepresented in “essential work,” more often requiring physical presence at a jobsite and direct client contact, thus contributing to the broader inequities in COVID-19 disease risk in Black and Hispanic populations.^{27–29}

4.1 | Study limitations

The purpose and strength of this study lie in comparing systems with different design characteristics, that is, an EHR system with individual demographic choices, solicited by clinical and nonclinical staff, compared to a new demographic form developed for this study that allowed self-report of multiple choices by the participants themselves.

A primary limitation of this study lies in the catchment of our clinics, which is influenced by state and local industry prevalence, existing relationships with specific employers, and the access to workers' compensation care enjoyed by particular industries or employers (and, therefore, by certain subpopulations). As such, our particular patient population influences the selection of study participants in a way that imposes limitations on the generalizability of our data and results.

The manual coding of industry, occupation, and injury data was a primary strength and limitation of this study. This coding adds significantly to occupational health data by discretely capturing work-specific characteristics that may define subpopulations with variable degrees of injury risk, the primary aim of this study. This coding was limited in that, only one industry, occupation, and injury (along the dimensions of OIICS coding) was captured for each new work injury, thus limiting the specificity of data collection in those cases where an individual works for multiple employers, in multiple codable occupations, or has multiple workers' compensation claims simultaneously. Encounter-based ICD-10 data was likely less susceptible to this effect than was occupation and industry.

Additionally, the NAICS and SOC codes derived from self-reported job title and employer name were not collected via the newly implemented demographic form and were not consistently collected via standard clinical processes. Typically, this information is available in clinical documentation recorded by the clinician, or on scanned patient intake forms, but sometimes one of or both of the job title or employer name are missing. In addition, the self-reported data may be entered with variations or abbreviations that may be difficult to decipher.

One further challenge is the low sample size for many linguistic and ethnic subgroups. Though the overall sample size is high, the small number of people of certain demographics, especially American Indians/Alaskan Natives and Native Hawaiians/Pacific Islanders means that potential trends may not be observed. This limitation can be overcome by ongoing surveillance of these data to develop a more robust sample size over time.

This study may be limited in generalizability as a different EHR software manufacturers or implementations within different health systems may capture demographic data in various ways, such as different categorization of race/ethnicity, or capturing multiple data elements (e.g., language) for a single demographic element. Coincident with this limitation is the implication that broader standardization of demographic data collection in clinical settings is an area of future study.

Another notable limitation of this study is that nonbinary gender identification was not included in the newly distributed demographic form.³⁰ This was due to the EHR not reliably collecting sexual orientation and gender identity (SOGI) data at the time of this project. Gender identification is increasingly understood as a determinant of health outcomes in general,³¹ and a factor which likely influences occupationally related health risks as well.³² Future work will require inclusion of both biological sex and gender identity demographic data to better characterize how these variables interact with other demographic, occupational, or environmental characteristics contributing to risk of injury or disease.

This project has demonstrated the ability to collect reliable data, allowing detection of disparities between various subpopulations and design of data-driven, community-based participatory research interventions that are relevant to those groups. To be effective and

meaningful, interventions must be specifically designed for at-risk communities with collaboration from and delivered in partnership with community groups that represent them, including employers.

Although data collection was disrupted due to the COVID-19 global pandemic, future demographic, industry, occupation, and injury data collection can be more closely integrated within the regular clinic workflow, administered by clerical staff with little modification required to existing processes. Collection will transition from use of the paper forms to tablet-based forms that import directly to the database with EHR data, without the need for coding by the research team. An electronic form will permit greater detail of collection via dropdown options based on responses, or by integrating race/ethnicities that arise during collection. For example, we plan to add Minnesota's Indigenous bands, tribes, and nations as subcategories under American Indian/Alaska Native. Finally, the form can be designed to provide immediate suggestions from any language and country in the world based on typed responses.

5 | CONCLUSION

Obtaining accurate and detailed patient demographic information is a critical step in addressing occupational health disparities. This study augmented existing EHR demographic data to develop meaningful sociocultural and occupationally specific data through the development and implementation of a novel demographic form. This study demonstrates that implementation of this form was feasible and that it provided a more comprehensive and accurate way to characterize important demographics such as race, ethnicity, language, educational attainment, and employment characteristics. Through this more nuanced approach, we will be able to identify specific groups (e.g., Hmong) that were not effectively identified in the EHR but may be disproportionately represented in more injury-prone occupational environments or beneficiaries of workplace safety interventions through existing community groups. Validation of this demographic form has been the successful first step to performing broader surveillance and design of focused interventions for the occupational health of our marginalized working populations.

ACKNOWLEDGMENT

This project was funded by the HealthPartners Institute, through the Regions Hospital Red Fund, St Paul, Minnesota (grant sponsor: Fozia Abrar, MD, MPH).

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

DISCLOSURE BY AJIM EDITOR OF RECORD

Paul A. Landsbergis declares that he has no conflict of interest in the review and publication decision regarding this article.

AUTHOR CONTRIBUTIONS

Andre G. Montoya-Barthelemy: conceptualization, major revisions, final approval, and accountability. **Karyn Leniek:** conceptualization, drafting and intellectual content, and funding acquisition. **Emily Bannister:** drafting, major revisions, funding acquisition, and supervision. **Marcus Rushing:** drafting, major revisions, and final approval. **Fozia A. Abrar:** conceptualization and final approval. **Tobias E. Baumann:** data management and analysis. **Madeleine Manly** and **Jonathan Wilhelm:** conceptualization, data management, and analysis. **Ashley Niece:** conceptualization, data management. **Scott Riester:** conceptualization, drafting, funding acquisition. **Hyun Kim:** conceptualization and data analysis. **Jonathan Sellman, Jay Desai, Paul J. Anderson, Ralph S. Bovard, and Nicolas P. Pronk:** major revisions and final approval. **Zeke J. McKinney:** drafting, supervision, data analysis, major revisions, and final approval. The corresponding author is responsible for ensuring that the descriptions are accurate and agreed by all authors.

ETHICS APPROVAL AND INFORMED CONSENT

This project was approved by the HealthPartners Institutional Review Board and written consent was obtained by the participants.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Andre G. Montoya-Barthelemy  <http://orcid.org/0000-0003-0567-279X>

Scott Riester  <http://orcid.org/0000-0002-5612-4956>

Hyun Kim  <http://orcid.org/0000-0002-5574-9532>

Zeke J. McKinney  <http://orcid.org/0000-0002-1858-8426>

REFERENCES

- Dickman SL, Himmelstein DU, Woolhandler S. Inequality and the health-care system in the USA. *Lancet*. 2017;389(10077):1431-1441. doi:10.1016/S0140-6736(17)30398-7
- Groenewold MR, Baron SL. The proportion of work-related emergency department visits not expected to be paid by workers' compensation: implications for occupational health surveillance, research, policy, and health equity. *Health Serv Res*. 2013;48(6 pt 1):1939-1959. doi:10.1111/1475-6773.12066
- Wuellner SE, Bonauto DK. Exploring the relationship between employer recordkeeping and underreporting in the BLS survey of occupational injuries and illnesses. *Am J Ind Med*. 2014;57(10):1133-1143. doi:10.1002/ajim.22350
- Fagan KM, Hodgson MJ. Under-recording of work-related injuries and illnesses: An OSHA priority. *J Safety Res*. 2017;60:79-83. doi:10.1016/j.jsr.2016.12.002
- Flynn MA, Eggerth DE, Jacobson CJ. Undocumented status as a social determinant of occupational safety and health: the workers' perspective. *Am J Ind Med*. 2015;58(11):1127-1137. doi:10.1002/ajim.22531
- Sabbath EL, Boden LI, Williams JAR, Hashimoto D, Hopcia K, Sorensen G. Obscured by administrative data? Racial disparities in occupational injury. *Scand J Work Environ Heal*. 2017;43(2):155-162. doi:10.5271/sjweh.3611
- Smith CK, Bonauto DK. Improving occupational health disparity research: testing a method to estimate race and ethnicity in a working population. *Am J Ind Med*. 2018;61(8):640-648. doi:10.1002/ajim.22850
- U.S. Bureau of Labor Statistics. Survey of occupational injuries and illnesses: concepts. Handbook of Methods. Updated 2020. Accessed May 29, 2019. <https://www.bls.gov/opub/hom/soii/concepts.htm>
- Elliott MN, Fremont A, Morrison PA, Pantoja P, Lurie N. A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. *Health Serv Res*. 2008;43(5 pt 1):1722-1736. doi:10.1111/j.1475-6773.2008.00854.x
- Andrews RM. Statewide hospital discharge data: collection, use, limitations, and improvements. *Health Serv Res*. 2015;50(S1):1273-1299. doi:10.1111/1475-6773.12343
- 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):719-723. doi:10.1007/s11606-014-3102-8
- 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.
- Riester SM, Leniek KL, Niece AD, et al. Occupational medicine clinical practice data reveal increased injury rates among Hispanic workers. *Am J Ind Med*. 2019;62(4):309-316. doi:10.1002/ajim.22949
- Horsky J, Drucker EA, Ramelson HZ. Accuracy and completeness of clinical coding using ICD-10 for ambulatory visits. *AMIA Annu Symp Proc*. 2017;2017:912-920.
- Schmitz M, Forst L. Industry and occupation in the electronic health record: an investigation of the National Institute for Occupational Safety and Health Industry and Occupation Computerized Coding System. *JMIR Med Informatics*. 2016;4(1):e5. doi:10.2196/medinform.4839
- Minnesota Compass. Immigration: overview. 2019. Accessed May 29, 2019. <https://www.mncompass.org/immigration/overview>
- Pfeifer ME, Sullivan J, Yang K, Yang W. Hmong population and demographic trends in the 2010 Census and 2010 American Community Survey. *Hmong Stud J*. 2012;13(2):1-31.
- Gambino CP, Trevelyan EN & Fitzwater JT Ten largest African-born countries of birth in the United States by selected metropolitan statistical areas: 2008-2012. 2014. Accessed June 1, 2019. https://www2.census.gov/library/publications/2014/acs/acsbr12-16_supptab2.pdf.
- Office for National Statistics Census. Household questionnaire—England. 2011. Accessed June 1, 2019. https://census.ukdataservice.ac.uk/media/50966/2011_english_household.pdf.
- Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42(2):377-381. doi:10.1016/j.jbi.2008.08.010
- Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. *J Biomed Inform*. 2019;95:103208. doi:10.1016/j.jbi.2019.103208
- McHugh ML. Interrater reliability: the kappa statistic. *Biochem Medica*. 2012;22(3):276-282. doi:10.11613/bm.2012.031
- Chan S, ed., Introduction: the Hmong Experience in Asia and the United States. *Hmong Means Free: Life in Laos and America*. Temple University Press; 1994:1-60.
- Okonofua BA. I am blacker than you. *SAGE Open*. 2013;3(3):215824401349916. doi:10.1177/2158244013499162
- Campbell S Racial and gender disparities within the direct care workforce: five key findings. 2017. Accessed June 22, 2019. <https://phinational.org/wp-content/uploads/2017/11/Racial-and-Gender-Disparities-in-DCW-PHI-2017.pdf>

26. Campbell S Workplace injuries and the direct care workforce. 2018. Accessed June 22, 2019. <https://phinational.org/wp-content/uploads/2018/04/Workplace-Injuries-and-DCW-PHI-2018.pdf>.
27. Lippert J, Rosing H, Tendick-Matesanz F. The health of restaurant work: a historical and social context to the occupational health of food service. *Am J Ind Med*. 2020;63(7):563-576. doi:10.1002/ajim.23112
28. Williams JC, Anderson N, Holloway T, Samford E, Eugene J, Isom J. Reopening the United States: Black and Hispanic workers are essential and expendable again. *Am J Public Health*. 2020;110(10):1506-1508. doi:10.2105/AJPH.2020.305879
29. Hughes MM, Groenewold MR, Lessem SE, et al. Update: characteristics of health care personnel with COVID-19—United States, February 12–July 16, 2020. *MMWR Morb Mortal Wkly Rep*. 2020;69(38):1364-1368. doi:10.15585/mmwr.mm6938a3
30. Deutsch MB, Green J, Keatley J, Mayer G, Hastings J, Hall AM. Electronic medical records and the transgender patient: recommendations from the World Professional Association for Transgender Health EMR Working Group. *J Am Med Inform Assoc*. 2013;20:700-703. doi:10.1136/amiajnl-2012-001472
31. Cahill S, Singal R, Grasso C, et al. Do ask, do tell: high levels of acceptability by patients of routine collection of sexual orientation and gender identity data in four diverse American community health centers. *PLOS One*. 2014;9(9):e107104.
32. Callahan EJ, Sitkin N, Ton H, Eidson-Ton S, Weckstein J, Latimore D. Introducing sexual orientation and gender identity into the electronic health record: one academic health center's experience. *Acad Med*. 2015;90:154-160. doi:10.1097/ACM.0000000000000467

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Montoya-Barthelemy AG, Leniek K, Bannister E, et al. Using advanced racial and ethnic identity demographics to improve surveillance of work-related conditions in an occupational clinic setting. *Am J Ind Med*. 2022;65:357-370. doi:10.1002/ajim.23332