Demographic Differences and Potential Bias From Automated Occupation Coding Among Mothers of Babies Born With or Without Cleft Lip and/or Cleft Palate in the Texas Birth Defects Registry

Omobola O. Oluwafemi, MPH, A. J. Agopian, PhD, Renata H. Benjamin, PhD, David Gimeno Ruiz de Porras, PhD, Charles J. Shumate, DrPH, and Jenil Patel, PhD

Objective: To compare maternal demographics based on occupation coding status and evaluate potential bias by excluding manually coded occupations. **Methods:** This case-control study assessed cases with clefts obtained from the Texas Birth Defects Registry. The NIOSH Industry and Occupation Computerized Coding System automatically coded occupations, with manual coding for unclassified cases. Maternal demographics were tabulated by occupation coding status (manual vs. automatic). Logistic regression examined associations between major occupation groups and clefts. **Results:** Automatic coding covered over 90% of all mothers. Building, grounds cleaning, and maintenance occupations, and office and administrative support occupations were significantly associated with cleft lip with or without cleft palate, even after excluding manually coded occupations. **Conclusion:** We found consistent associations before and after excluding manually coded data for most comparisons, suggesting that machine learning can facilitate occupation-related birth defects research.

Keywords: cleft lip, cleft palate, birth defects, NIOCCS, occupation

LEARNING OUTCOMES

- Compare the demographic characteristics of mothers whose occupations are automatically coded by software to mothers whose occupations need to be manually coded.
- Evaluate potential bias in odds ratios when mothers whose occupations were not automatically coded by software are excluded from analyses.
- Discuss the benefits of using automatic coding software to code occupational data for birth defects research, which often requires very large sample sizes due to the rarity of the outcome.

Several previous studies have assessed maternal and paternal occupations or specific occupational exposures and their relationships with the occurrence of birth defects.[1–5] Most often, such studies use free-text descriptions of maternal/paternal occupations from survey responses[2–4] or from birth certificate data that are available through statewide surveillance systems.[6,7] As the commonly used industry and occupation terms are very detailed, the data obtained from free-text responses must be systematically classified by postcollection processing that assigns standard occupational codes. For example, in the National Birth Defects Prevention Study (NBDPS), trained National Institute for Occupational Safety and Health (NIOSH) industrial hygienists or abstractors manually assigned Standard Occupation Classification (SOC) codes and North American Industry Classification System (NAICS) codes to each job based on text descriptions of the occupation.[7,8] Any discrepancies are then resolved by consulting an additional rater.[7,8] Similarly, the European Registration of Congenital Anomalies (EUROCAT) surveillance system codes occupations and industrial activities with the help of an industrial hygiene expert who interprets the job descriptions provided and assigns occupational and industrial codes accordingly.[9,10] Complicated job descriptions require decisions by consensus among raters.[9]

While manual coding of occupation categories is considered the gold standard method, it can be cost prohibitive to perform manual coding in large datasets (eg, >10,000 subjects). Thus, semiautomated approaches have been implemented using standardized software such as the NIOSH Industry and Occupation Computerized Coding System (NIOCCS).[11] The NIOCCS is a publicly available software that translates industry and occupation (I&O) text found in survey, birth or death certificates, and health records into standardized I&O codes. The software is periodically updated and improved; it is currently on its fourth version, which now incorporates machine learning into occupational code assignment.[11] The NIOCCS automated coding algorithm greatly accelerates the process of categorizing free-text occupational data and has previously been shown to be up to 90% accurate.[12] However, some degree of supplemental manual coding for a proportion

From the Department of Epidemiology, Human Genetics and Environmental Sciences, UTHealth Houston School of Public Health, Houston, Texas (O.O.O., A.J.A., R.H.B.); Department of Epidemiology, Human Genetics and Environmental Sciences, UTHealth Houston School of Public Health, San Antonio, Texas (D.G.R.); Birth Defects Epidemiology and Surveillance Branch, Texas Department of State Health Services, Austin, Texas, (C.J.S.); and Department of Epidemiology, Human Genetics and Environmental Sciences, UTHealth Houston School of Public Health, Dallas, Texas (J.P.).

Funding sources: This work was supported by T42 OH008421/OH/NIOSH CDC HHS/United States training grant and funding from the Centers for Disease Control and Prevention (CDC) as part of a birth defects surveillance cooperative agreement with the Texas Department of State Health Services (TX DSHS) (NU50DD000102). The TX DSHS Birth Defects Epidemiology and Surveillance Branch is supported in part by Maternal and Child Health Services Title V Block Grant funds from the Health Resources and Services Administration (HRSA). The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement by, CDC, TX DSHS, or HRSA.

Conflicts of interest: None declared.

Specific author contributions: O.O.O. conducted data analysis and data interpretation, drafted the original manuscript, and assisted with manuscript revision. A.J.A. and R.H.B. contributed to data analysis, data interpretation, and manuscript revision. D.G.R.D.P. and C.J.S. assisted with data interpretation and manuscript revision. J.R.P. supervised all research and assisted with manuscript revision.

Data availability statement: Due to data confidentiality restrictions and existing data agreements, these data cannot be shared.

EQUATER Network Reporting Guidelines: We adhered to the STROBE reporting guidelines for this case-control study.

Ethical considerations and disclosure(s): Our protocol was approved by the UTHealth Institutional Review Board. The Texas Birth Defects Registry has legislative authority to collect data without informed consent.

Supplemental digital contents are available for this article. Direct URL citation appears in the printed text and is provided in the HTML and PDF versions of this article on the journal's Web site (www.joem.org).

Address correspondence to: Jenil R. Patel, PhD, Department of Epidemiology, Human Genetics and Environmental Sciences, UTHealth Houston School of Public Health, 2777 N Stemmons Fwy, Suite 8400, Dallas, TX 75207 (Jenil.Patel@uth.tmc.edu).

DOI: 10.1097/JOM.0000000000003189

(eg, up to ~15.3%)[7] of data has still typically been required (eg, select free-text responses that cannot be automatically coded by the software). This manual coding can be cumbersome, costly, time-consuming, and potentially error prone,[13] though less costly and time-consuming compared to a full manual review of all subjects.

Thus, it is tempting to consider the added value of incorporating data from manually reviewed subjects compared to the possibility of simply restricting analyses to only subjects with successful automated occupational assignment. However, few studies have assessed the effectiveness and potential differences in bias between automatically and manually coded occupations. One such study that compared differences in automatic and manual coding focused on accuracy of occupational exposures assessment, specifically.[14] Another study measured disagreement between manual occupation coding and NIOCCS automatic coding, but it did not compare the demographic profiles or occupation profiles of those that were manually coded compared to automatically coded.[15]

Thus, the use of state vital records data to advance our knowledge of occupational exposures has been hampered, not by lack of data on occupations but by data management (ie, translating free text on birth certificates to standard occupation categories). Consequently, our relatively limited understanding of occupational risk factors is due, in part, to this limitation in data processing.

Learning whether individuals with manually coded occupations differ from those with automatically coded occupations might help researchers to make decisions for future studies about whether involving manual coding as a second classification step is still necessary with the current version of NIOCCS, or if automatically coded subjects might be sufficient for analysis when resources are limited for supplemental manual review. To better understand the potential impact of such manual review, we used data from the Texas Birth Defects Registry (TBDR) to follow up on a previous association study on maternal occupations and orofacial clefts,[7] to (i) compare the demographic characteristics and distributions of occupations of mothers of babies born with or without cleft lip and/or cleft palate on the basis of automatically versus manually coded occupations in Texas, and (ii) quantify estimates of potential bias (if any) in measures of association resulting from exclusion of manually coded occupations.

METHODS

Study Population

We analyzed data from the Texas Birth Defects Registry (TBDR), one of the largest birth defects registries in the United States. The TBDR is a population-based, active surveillance registry that conducts birth defects surveillance among all deliveries to women residing in Texas and is maintained by the Texas Department of State Health Services (TDSHS) Birth Defects Epidemiology and Surveillance Branch.[16] To be included in the registry, potential cases must have a documented structural birth defect or chromosomal anomaly that is diagnosed no later than 1 year after delivery. Data on potential cases are abstracted from medical records, and confirmed birth defects are coded using a modified six-digit British Pediatric Association (BPA) Classification of Diseases. These data are linked to sociodemographic data from birth certificates or fetal death certificates managed by the Texas Vital Statistics Unit. Birth certificate data were also used to determine maternal occupation as well as to obtain confounders of interest (described below).

In our previously published study, data were obtained from the TBDR among deliveries from January 1, 1999, to December 31, 2009, and included cases with cleft lip with or without cleft palate (CLP) (BPA codes: 749.100 to 749.220) or cleft palate only (CP) (749.000 to 749.090) (n = 7031).[7] To limit heterogeneity, subjects that were identified as having a syndrome, chromosomal abnormality, additional major birth defect present, or malformation complex were excluded,[17] which resulted in 4249 isolated cases. Additionally, analyses were restricted to liveborn individuals (n = 4207), as maternal occupation data were not available for fetal deaths (n = 42). The unmatched control group consisted of 6000 randomly selected live births without any birth defect delivered in Texas during the study period. The current analysis utilized the same study population.

Our protocol was approved by the UTHealth Institutional Review Board. The Texas Birth Defects Registry has legislative authority to collect data without informed consent. The study methods adhered to the Strengthening the Reporting of Observational studies in Epidemiology (STROBE) checklist for cohort studies (Supplemental Digital Content, http://links.lww.com/JOM/B658).

Occupation Classification and Coding

Prior Classification and Coding

We previously conducted a study[7] using the NIOCCS, which did not compare subjects with automatically versus manually assigned occupations. Additional methods to classify occupation for this dataset have been previously described[7]; briefly, data on maternal industry, occupation, age, and education were input into the program. After adjusting the software settings, the program output file included 2010 Standard Occupation Classification (SOC) codes (http://www.bls.gov/soc/). Subjects that were not coded by the program were manually coded by an abstractor trained in NIOSH occupation coding. Mothers without an occupation code or mothers whose occupation code corresponded to student, housewife, or unemployed were excluded in order to limit healthy worker bias.[7,18,19]

The final categorized maternal occupations were classified under 1 of 23 major groups, according to the first two digits of the SOC code. These occupations were further classified into six high-level aggregation groups. And finally, the high-level aggregation occupations were classified into two broad occupation categories.[20] The first broad category (group 1) included the first SOC high-level aggregation group, which included occupations related to management, business, science, and arts. The second broad category (group 2) included SOC high-level aggregation groups 2–6 and included occupations such as service, sales, construction, or military-specific occupations.

Current Classification and Coding

This current analysis repeated the abovementioned classification and coding methods utilizing the newest version of NIOCCS, which now incorporates machine learning. Subjects that were not coded by the program were manually coded by an abstractor (O.O.O.), and mothers without an occupation code or mothers whose occupation code corresponded to student, housewife, or unemployed were excluded.

Statistical Analysis

Counts and frequencies were tabulated for demographic characteristics of control mothers by whether occupations were automatically coded or manually coded. These values were compared using χ^2 tests. Counts and frequencies were also tabulated for each occupation category of control mothers whose occupations were automatically coded versus manually coded, and these values were compared using χ^2 tests. Additionally, counts and frequencies were tabulated for demographic characteristics of mothers of cases with (1) CLP and (2) CP by occupation coding status (separately for automatically coded vs. manually coded subjects).

For the full group (automatically coded and manually coded occupations combined), we conducted logistic regression analyses to evaluate the association between each of the 23 major occupation categories and CLP and CP separately. We also stratified the analyses by occupation coding status (automatically coded vs. manually coded) and repeated these logistic regression analyses within each stratum.

The reference group for each comparison was the total of all subjects in all other major occupation groups (eg, management occupations versus all nonmanagement occupations for the assessment of management occupations). Additionally, logistic regression analysis was conducted between the two broad occupation groups (defined above), with group 2 serving as the reference group. Both unadjusted and adjusted logistic regression models were conducted. Multivariable models were adjusted for maternal age at delivery, education, race/ethnicity, parity, any diabetes, and smoking during pregnancy.

To better understand the potential impact of excluding subjects with manually coded occupations using an applied association analysis example, we calculated the potential percentage bias induced by restricting to the subgroup of subjects with automatically assigned occupations compared to using the full group (auto coded + manually coded). Specifically, we estimated the change in odds ratios (ORs) for cleft risk between these groups using the following formula: [(auto-coded OR − full OR)/full OR]. The bias was calculated for both the crude and adjusted CLP ORs as well as the crude and adjusted CP ORs.

Across all analyses, comparisons involving cells with less than five individuals in an occupation group were not assessed. Analyses were conducted using SAS, version 9.4.[21]

RESULTS

Among 10,207 initial subjects, the NIOCCS program automatically coded 9335 (91.5%) individuals in our dataset, including students, housewives, and other nonworking individuals. We manually assigned occupation codes to 287 out of the 872 individuals whose occupations were not automatically coded by NIOCCS, and the remaining 585 participants were missing occupation information. After excluding 6342 (62.1%) nonworking mothers and mothers with missing occupation information, there were a total of 3865 subjects analyzed (1063 cases with CLP, 511 cases with CP, and 2291 controls). Demographic characteristics of control mothers are shown in Table 1. Of the 2291 control mothers, occupations were automatically coded for 2122 mothers and manually coded for 169 mothers (7.4%). There was a significant demographic difference between automatically and manually coded control mothers by race/ethnicity ($P = 0.001$), marital status ($P = 0.006$), and diabetes status ($P = 0.009$). More precisely, manually coded control mothers were more likely to be Black (20.1%) or Hispanic (42.0%) compared to automatically coded control mothers (49.0% White, 12.6% Black, and 33.4% Hispanic). About 70% of control mothers that were automatically coded were married compared to about 60% of manually coded control mothers. Differences were not observed by maternal age at delivery, maternal BMI, maternal education, parity, and smoking.

Table 2 displays the occupational groups of control mothers stratified by automatic versus manual coding status. Among the control mothers, 93.5% of mothers in group 1 were automatically coded and 92.1% of mothers in group 2 were automatically coded. There was no overall significant difference between manually versus automatically coded occupations in group 1 versus group 2. There were significant differences between automatically and manually coded

TABLE 1. Demographic Characteristics of Control Mothers Whose Occupations Were Automatically Coded and Manually Coded and Who Delivered in Texas, 1999–2009

	Automatically Coded (*n* = 2122)	Manually Coded (*n* = 169)	
Demographic Characteristics	***n* (%)**	***n* (%)**	***P*[a]**
Maternal age at delivery, y			0.580
<20	111 (5.2)	12 (7.1)	
20–34	1723 (81.2)	135 (79.9)	
≥35			
Maternal BMI,[b] kg/m^2			0.928
Underweight (<18.5)	27 (3.0)	NR[c]	
Normal (18.5 to <25)	399 (44.8)	NR	
Overweight (25 to <30)	252 (28.3)	NR	
Obese (30+)	212 (23.8)	NR	
Maternal education			0.305
<High school	217 (10.2)	23 (13.6)	
High school	646 (30.4)	52 (30.8)	
≥College	1248 (58.8)	91 (53.9)	
Maternal race/ethnicity			0.001
White	1039 (49.0)	57 (33.7)	
Black	268 (12.6)	34 (20.1)	
Hispanic	709 (33.4)	72 (42.0)	
Others	101 (4.8)	7 (4.1)	
Mother married			0.006
Yes	1483 (69.9)	101 (59.8)	
No	639 (30.1)	68 (40.2)	
Parity			0.342
0	881 (41.5)	63 (37.3)	
≥1	1206 (56.8)	101 (59.8)	
Any diabetes			0.009
Yes	82 (3.9)	NR	
No	2040 (96.1)	NR	
Smoking during pregnancy			0.374
Yes	139 (6.6)	14 (8.3)[d]	
No	1969 (92.8)	153 (90.5)	

BMI indicates body mass index; NR, not reported.

[a] *P* values were calculated using χ^2 tests.

[b] Maternal BMI available only during 2005–2009 (total *n* = 959 among controls).

[c] NR due to small cell counts.

[d] Percentage may not sum to 100% due to missing data.

TABLE 2. Automatically Coded and Manually Coded Occupations of Control Mothers Who Delivered in Texas, 1999–2009 (n = 2291)

	Automatically Coded (n = 2122)	Manually Coded (n = 169)	
Occupation group	***n* (%)**	***n* (%)**	***P*[a]**
Broad occupation group			
Group 1[b]	870 (93.5)[c]	61 (6.5)	0.212[d]
Group 2[e]	1252 (92.1)	108 (7.9)	
Major groups			
Group 1			
Management occupations	137 (89.0)	17 (11.0)	0.072[f]
Business and financial operations occupations	136 (88.3)7	18 (11.7)	0.034
Computer and mathematical science occupations	33 (86.8)	5 (13.2)	0.169
Healthcare practitioner and technical occupations	192 (96.0)	8 (4.0)	0.056
Group 2			
Healthcare support occupations	115 (92.7)	9 (7.3)	0.959
Food preparation and serving-related occupations	105 (95.5)	5 (4.5)	0.244
Sales and related occupations	311 (95.1)	16 (4.9)	0.063
Office and administrative support occupations	448 (88.2)	60 (11.8)	<0.001
Transportation and material moving occupations	33 (78.6)	9 (21.4)	<0.001

Occupational groups with cells less than five were not presented due to data use requirements.

[a] P value was calculated using χ^2 tests.

[b] Included occupations related to management; business and financial operations; computer and mathematical science; architecture and engineering; life, physical, and social science; community and social service; legal work; education, training, and library; arts, design, entertainment, sports, and media; and healthcare practitioner and technical operations.

[c] Row percentages are presented.

[d] P value comparing high level group 1 to high level group 2.

[e] Included occupations related to healthcare support; protective service; food preparation and serving; building and grounds cleaning and maintenance; personal care and service; sales; office and administrative support; farming, fishing, and forestry; construction and extraction; installation, maintenance, and repair; production; transportation and material moving; and armed forces.

[f] The reference group for each χ^2 comparison was the total of all subjects in the other major occupation groups.

business and financial operations occupations (P = 0.034). Additionally, there were significant differences between both occupation coding statuses for office and administrative support occupations (P < 0.001) and transportation and material moving occupations (P < 0.001), but no significant differences for six other occupational groups were assessed. For comparison, supplemental Table S1 (http://links.lww.com/JOM/B659) shows the occupational groups of mothers of cases stratified by automatic versus manual coding status. For mothers of cases with CLP, there were significant differences between both occupation coding statuses for group 1 compared to group 2 (P = 0.043), education, training, and library occupations compared to all other occupations (P = 0.018), and office and administrative support occupations (P < 0.001) compared to all other occupations, but no significant differences for 18 other occupation groups. Among mothers of cases with CP, there was a significant difference between both occupation coding groups among healthcare support occupations (P = 0.049) and office and administrative support occupations (P = 0.012) compared to all other occupations, but no significant differences for 18 other occupation groups.

Demographic characteristics and occupation coding status of mothers of cases with oral clefts are shown in Table 3. Among cases with CLP, 986 mothers had occupations that were automatically coded, and 77 (7.2%) mothers had occupations that were manually coded. There were 470 mothers of cases with CP that had occupations that were automatically coded compared to 41 (8.0%) mothers of cases with CP whose occupations were manually coded. Among cases with CLP, mothers that were automatically coded differed significantly from mothers that were manually coded in regard to maternal age at delivery (P = 0.020), maternal education (P = 0.013), maternal race/ethnicity (P = 0.021), and marital status (P = 0.007). A greater proportion of mothers whose occupations were manually coded were <20 years old (11.7%) compared to mothers whose occupations were manually coded (4.6%). Mothers whose occupations were manually coded were more likely to have completed high school or less education (52.7%) compared to mothers whose occupations were automatically coded (41.6%). Among married mothers, 72.9% had their occupations automatically coded compared to 58.4% who had their occupations manually coded. Among cases with CP, there were no significant differences in demographic characteristics among automatically coded mothers compared to manually coded mothers.

Table 4 displays the adjusted odds ratios (aORs) with 95% confidence intervals (CIs) for the association between each maternal occupation and occurrence of CLP in offspring using the subsets of subjects with (1) automatic coding and (2) the combination of both manual and automatic coding. The bias quantifying the estimated impact of excluding subjects with occupations that could not be automatically coded (the ratio of ORs for CLP risk among automatically coded occupations to the combination of automatically and manually coded occupations) was <0.10 for most occupations, with the exceptions of 0.159 for transportation and material moving occupations and 0.145 for architecture and engineering occupations. Among the full group with combined automatic + manual coding, there was a significant association between building and grounds cleaning and maintenance occupations and CLP (aOR: 2.21, 95% CI: 1.30, 3.76) and office and administrative support occupations and CLP (aOR: 0.77, 95% CI: 0.63, 0.93) after adjustment for maternal age at delivery, education, race/ethnicity, parity, any diabetes, and smoking during pregnancy, and these associations remained significant upon exclusion of the manually coded occupation group.

These analyses were repeated for CP (Table 5). The magnitude of the bias estimate comparing the ORs among automatically coded subjects to the ORs among combined manually and automatically coded subjects was >0.10 for 6 of the 23 comparisons, most of which were not significantly associated in either comparison, and for which the 95% CIs overlapped substantially between each of the two groups. Of note, the bias estimate for architecture and engineering occupations was 0.205 with associations remaining significant in both groups; however, this association was attenuated in the full group (aOR: 2.64, 95% CI: 1.27, 5.49) compared to the subset with automatically coded subjects. Among the full group, CP was also significantly associated with office and administrative support occupations (aOR: 0.78, 95% CI: 0.60, 1.00), and these associations remained significant when

TABLE 3. Demographic Characteristics and Occupation Coding Status of Mothers of Cases With Oral Clefts Delivered in Texas, 1999–2009

	CLP Cases			CP Cases		
	Automatically Coded (*n* = 986)	Manually Coded (*n* = 77)		Automatically Coded (*n* = 470)	Manually Coded (*n* = 41)	
Demographic Characteristics	*n* (%)	*n* (%)	*P*[a]	*n* (%)	*n* (%)	*P*[a]
Maternal age at delivery, y			0.020			0.746
<20	45 (4.6)	9 (11.7)		19 (4.0)	NR[b]	
20–34	799 (81.0)	56 (72.7)		367 (78.1)	NR	
≥35	142 (14.4)	12 (15.6)		84 (17.9)	NR	
Maternal BMI,[c] kg/m^2			0.347			0.462
Underweight (<18.5)	19 (4.4)	NR		10 (4.7)	NR	
Normal (18.5 to <25)	209 (48.1)	NR		84 (39.4)	NR	
Overweight (25 to <30)	104 (23.9)	NR		61 (28.6)	NR	
Obese (30+)	103 (23.7)	NR		58 (27.2)	NR	
Maternal education			0.013			0.161
<High school	113 (11.5)	5 (6.6)		60 (12.8)	NR	
High school	296 (30.1)	35 (46.1)		143 (30.4)	NR	
≥College	573 (58.4)	36 (47.4)		267 (56.8)	NR	
Maternal race/ethnicity			0.021			0.205
White	538 (54.7)	NR		261 (55.5)	17 (41.5)	
Black	74 (7.5)	NR		55 (11.7)	5 (12.2)	
Hispanic	326 (33.1)	NR		127 (27.0)	14 (34.2)	
Others	46 (4.7)	NR		27 (5.8)	5 (12.2)	
Mother married			0.007			0.218
Yes	718 (72.9)	45 (58.4)		330 (70.2)	25 (61.0)	
No	267 (27.1)	32 (41.6)		140 (29.8)	16 (39.0)	
Parity			0.994			0.921
0	388 (40.0)	30 (40.0)		175 (38.3)	15 (37.5)	
≥1	581 (60.0)	45 (60.0)		282 (61.7)	25 (62.5)	
Any diabetes			0.260			0.281
Yes	48 (4.9)	6 (7.8)		18 (3.8)	NR	
No	938 (95.1)	71 (92.2)		452 (96.2)	NR	
Smoking during pregnancy			0.552			0.081
Yes	88 (9.0)	5 (6.9)		52 (11.1)	NR	
No	889 (91.0)	67 (93.1)		416 (88.9)	NR	

BMI indicates body mass index; CLP, cleft lip with or without cleft palate; CP, cleft palate only.

[a] *P* values were calculated using χ^2 tests.

[b] NR due to small cell counts.

[c] Maternal BMI available only during 2005–2009.

manually coded occupations were excluded. Among only automatically coded mothers, there were marginally significant associations between broad CP and group 1 (aOR: 1.26, 95% CI: 1.00, 1.59) and business and financial operations occupations (aOR: 1.46, 95% CI: 1.01, 2.10), which were not significant among the full group.

The bias estimates for the crude and adjusted ORs for CP and CLP were noted in Supplemental Table S2 (http://links.lww.com/JOM/B659). For most ORs, there was no substantial difference between the bias estimates for the crude and adjusted ORs. Among mothers of CLP cases, the crude bias estimate for healthcare practitioner and technical occupations was 0.013, but it changed direction to −0.032 for the adjusted estimate. Among mothers of CP cases, the crude bias estimate for education, training, and library occupations was −0.013, and the adjusted bias estimate was 0.009. Additionally, the crude bias estimate for transportation and material moving occupations was 0.049, and the adjusted bias estimate was −0.026.

DISCUSSION

In the first study of its kind focusing on comparing the performance of automatic and manual coding of occupations for studies on risk assessment for birth defects, we observed a few demographic differences between individuals that were manually and automatically coded by NIOCCS, although the occupation profiles associated with orofacial clefts were fairly similar when comparing automatic to automatic + manual codes, which suggests that automatic coding through NIOCCS may be sufficient for at least certain occupation-related analyses. Furthermore, the incorporation of machine learning algorithms in NIOCCS from 2021 is likely to bring more refinements and a potential increase in the number of automatically coded occupations in the future. The demographic differences were particularly noted with respect to key factors such as race/ethnicity and smoking, which are suspected risk factors for orofacial clefts, among other birth defects, though these differences did not seem to translate into major differences in the associated occupations. Our findings were focused on birth defects research, where overall only a few studies have assessed occupations or occupational exposures as risk factors, owing to rare and low prevalence of outcomes.[2–4,22] In this study, we focused only on cases with CLP/CP, since a previous assessment using this dataset had been conducted.[7] This assisted us in comparing our current estimates to the estimates of the previous analysis, which used a prior version of NIOCCS without the machine learning component. We were able to assess the reliability of the updated NIOCCS program by comparing our current results to our prior results, which had a greater proportion of manually coded participants. Obtaining similar results between both studies would indicate that the automatically coded

TABLE 4. Odds Ratios and Percentage Bias for the Comparison of Each Maternal Occupation and CLP Among Automatically Coded and Manually Coded Occupations in Texas, 1999–2009

	Automatically Coded Subset	Full Group	
Occupation Group[a]	Adjusted OR[b] (95% CI)	Adjusted OR[b] (95% CI)	CLP Adjusted Bias Estimate[c]
Broad occupation group			
Group 1[d]	1.17 (0.98, 1.39)	1.15 (0.97, 1.36)	0.017
Group 2[e]	1.00	1.00	
Major groups			
Group 1			
Management occupations	0.97 (0.71, 1.32)	0.96 (0.71, 1.29)	0.010
Business and financial operations occupations	1.16 (0.86, 1.57)	1.15 (0.86, 1.52)	0.009
Computer and mathematical science occupations	1.47 (0.84, 2.59)	1.40 (0.82, 2.39)	0.050
Architecture and engineering occupations	0.95 (0.41, 2.21)	0.83 (0.37, 1.90)	0.145
Life, physical, and social science occupations	0.44 (0.18, 1.09)	0.41 (0.17, 1.01)	0.073
Community and social service occupations	1.24 (0.69, 2.21)	1.26 (0.73, 2.19)	−0.016
Legal occupations	1.40 (0.76, 2.61)	1.41 (0.76, 2.62)	−0.007
Education, training, and library occupations	0.95 (0.74, 1.22)	0.96 (0.75, 1.23)	−0.010
Arts, design, entertainment, sports, and media occupations	1.22 (0.68, 2.20)	1.15 (0.64, 2.06)	0.061
Healthcare practitioner and technical occupations	1.23 (0.95, 1.59)	1.27 (0.94, 1.57)	−0.032
Group 2			
Healthcare support occupations	1.20 (0.86, 1.66)	1.27 (0.93, 1.73)	−0.055
Protective service occupations	0.87 (0.36, 2.09)	0.80 (0.34, 1.92)	0.088
Food preparation and serving-related occupations	0.92 (0.64, 1.32)	0.92 (0.64, 1.31)	0.000
Building and grounds cleaning and maintenance occupations	2.24 (1.32, 3.83)	2.21 (1.30, 3.76)	0.014
Personal care and service occupations	1.12 (0.80, 1.58)	1.13 (0.80, 1.58)	−0.009
Sales and related occupations	0.91 (0.72, 1.13)	0.91 (0.73, 1.14)	0.000
Office and administrative support occupations	0.74 (0.60, 0.90)	0.77 (0.63, 0.93)	−0.039
Production occupations	1.05 (0.65, 1.72)	1.08 (0.68, 1.74)	−0.028
Transportation and material moving occupations	1.31 (0.74, 2.31)	1.13 (0.67, 1.92)	0.159
Armed forces	0.42 (0.16, 1.10)	0.42 (0.16, 1.10)	0.000

CLP indicates cleft lip with or without cleft palate; OR, odds ratio; CI, confidence interval.

[a] Reference group for each comparison was the total of all subjects in the other major occupation groups.

[b] Adjusted for maternal age at delivery, education, race/ethnicity, parity, any diabetes, and smoking during pregnancy.

[c] Bias estimate quantifying the ratio of ORs for cleft risk among the subset with only automatically coded occupations versus the combination of automatically and manually coded occupations in Texas, 1999–2009. Bias calculated as [(auto-coded − full group)/full group].

[d] Included occupations related to management; business and financial operations; computer and mathematical science; architecture and engineering; life, physical, and social science; community and social service; legal work; education, training, and library; arts, design, entertainment, sports, and media; and healthcare practitioner and technical operations.

[e] Included occupations related to healthcare support; protective service; food preparation and serving; building and grounds cleaning and maintenance; personal care and service; sales; office and administrative support; farming, fishing, and forestry; construction and extraction; installation, maintenance, and repair; production; transportation and material moving; and armed forces.

samples in the current version of NIOCCS were in alignment with the manually coded samples of our prior study. In comparison to this prior study, which had 84.7% of samples automatically coded by NIOCCS, the updated version of NIOCCS automatically coded 91.5% of our dataset.[7]

Regarding maternal occupation and CLP, there were certain occupations for which we found significantly increased overall odds ratios for the full group. These were for building and grounds cleaning and maintenance occupations as well as office and administrative support occupations. The cleaning and custodial occupation, in general, performs a wide array of tasks (sweeping, waxing, disinfecting) and is likely to be exposed to occupational hazards such as cleaning agents and chemicals.[2,23,24] In 2021, there were over 2.2 million reported cleaning jobs in the United States, which supports the need to address occupational hazards in this sizeable workforce.[25] Among cases with CP, there were positive associations with architecture and engineering occupations among the full group. This finding is consistent with the prior study conducted on this dataset that also found significantly associated results for this occupational group and clefts.[7] In architectural fields, some women are exposed to organic solvents, which have been shown to be associated with orofacial clefts, which may contribute to the observed association.[7,24] We also observed negative associations with CP within the full group of office and administrative support occupations. These results are consistent with our prior study conducted on this dataset, which found a marginally significant association.[7]

We sought to explore the demographic characteristics of automatically versus manually coded participants as part of the overall comparison of the two groups. Among control mothers, maternal race/ethnicity, marital status, and diabetes status were significantly different between automatically versus manually coded mothers. Among mothers of cases with CLP, there were demographic differences between automatically and manually coded mothers regarding maternal age, education, race/ethnicity, marital status, and smoking status. These results indicate that excluding participants that are not automatically coded by NIOCCS would potentially exclude individuals that are more likely to be younger, of Black or Hispanic race/ethnicity, unmarried, and holding a high school degree or less. Such exclusions could lead to disparities in occupational research. These differences may also suggest that a simplistic approach of not conducting manual coding and restricting analyses to those with only automated coded occupations may result in skewing the distribution of population characteristics, which could be especially problematic in association analyses if the characteristics were associated with the exposure and outcome. However, the association analysis did not reveal major differences while comparing automatic and manual versus automatic only.

Beyond birth defects research, NIOCCS may be helpful in other health contexts, particularly as it relates to work-related illness or injury. Including I&O information in electronic health records (EHR), for example, could provide healthcare professionals and healthcare systems with information regarding workplace exposures or conditions that

TABLE 5. Odds Ratios and Percentage Bias for the Comparison of Each Maternal Occupation and CP Among Automatically Coded and Manually Coded Occupations in Texas, 1999–2009

	Automatically Coded Subset	Full Group	
Occupation Group[a]	**Adjusted OR**[b] **(95% CI)**	**Adjusted OR**[b] **(95% CI)**	**CP Adjusted Bias Estimate**[c]
Broad occupation group			
Group 1[d]	1.26 (1.00, 1.59)	1.16 (0.93, 1.45)	0.086
Group 2[e]	1.00	1.00	
Major groups			
Group 1			
Management occupations	0.94 (0.62, 1.43)	0.94 (0.63, 1.39)	0.000
Business and financial operations occupations	1.46 (1.01, 2.10)	1.38 (0.97, 1.95)	0.058
Computer and mathematical science occupations	0.96 (0.41, 2.21)	0.91 (0.42, 1.98)	0.055
Architecture and engineering occupations	3.18 (1.49, 6.77)	2.64 (1.27, 5.49)	0.205
Life, physical, and social science occupations	1.25 (0.55, 2.80)	1.31 (0.61, 2.83)	−0.046
Community and social service occupations	1.15 (0.53, 2.53)	1.02 (0.47, 2.23)	0.128
Legal occupations	1.06 (0.43, 2.61)	1.21 (0.52, 2.83)	−0.124
Education, training, and library occupations	1.10 (0.79, 1.52)	1.09 (0.79, 1.50)	0.009
Arts, design, entertainment, sports, and media occupations	1.54 (0.74, 3.20)	1.40 (0.68, 2.90)	0.100
Healthcare practitioner and technical occupations	0.78 (0.53, 1.14)	0.71 (0.48, 1.05)	0.099
Group 2			0.000
Healthcare support occupations	0.89 (0.55, 1.42)	1.01 (0.66, 1.56)	−0.119
Food preparation and serving-related occupations	1.04 (0.66, 1.66)	1.02 (0.65, 1.61)	0.020
Building and grounds cleaning and maintenance occupations	1.19 (0.52, 2.72)	1.18 (0.52, 2.67)	0.008
Personal care and service occupations	0.95 (0.59, 1.52)	0.95 (0.59, 1.51)	0.000
Sales and related occupations	1.16 (0.87, 1.53)	1.16 (0.88, 1.52)	0.000
Office and administrative support occupations	0.75 (0.57, 0.99)	0.78 (0.60, 1.00)	−0.039
Production occupations	0.91 (0.46, 1.79)	1.06 (0.56, 1.98)	−0.142
Transportation and material moving occupations	1.14 (0.53, 2.43)	1.17 (0.59, 2.30)	−0.026

CP indicates cleft palate; OR, odds ratio; CI, confidence interval.

[a] Reference group for each comparison was the total of all subjects in the other major occupation groups.

[b] Adjusted for maternal age at delivery, education, race/ethnicity, parity, any diabetes, and smoking during pregnancy.

[c] Bias estimate quantifying the ratio of ORs for cleft risk among the subset with only automatically coded occupations versus the combination of automatically and manually coded occupations in Texas, 1999–2009. Bias calculated as [(auto-coded − full group)/full group].

[d] Included occupations related to management; business and financial operations; computer and mathematical science; architecture and engineering; life, physical, and social science; community and social service; legal work; education, training, and library; arts, design, entertainment, sports, and media; and healthcare practitioner and technical operations.

[e] Included occupations related to healthcare support; protective service; food preparation and serving; building and grounds cleaning and maintenance; personal care and service; sales; office and administrative support; farming, fishing, and forestry; construction and extraction; installation, maintenance, and repair; production; transportation, and material moving; and armed forces.

contribute to decreased health and quality of life.[26] Incorporating relevant I&O information into EHR would also allow for population-level intervention efforts and improved surveillance of occupation-related injuries or illnesses.[26,27]

At an individual level, access to workplace information can allow healthcare professionals to quickly identify potential workplace exposures or circumstances (eg, chemical, psychosocial, or physical hazards) that may be compromising a patient's health. At the organization level, NIOCCS can be used to efficiently include I&O information in health records, which can then be used to implement and evaluate various services such as healthcare coverage, access and utilization, implementation of workplace health and wellness programs, and workers' compensation costs.[26] At the population level, such information can be used for surveillance measures as well as prevention measures or health equity assessments.[27] For example, compared to other demographic groups, Black women are disproportionally overrepresented in low-wage healthcare jobs, which are characterized by a lack of benefits and dangerous working conditions.[28]

While our findings focus on a single research question and different conclusions may have been made for a research question related to a different outcome, our findings seem to indicate that the automatic coding algorithm by NIOCCS may be sufficient for occupational research. Further, the quality of the input dataset can greatly affect the accuracy of automatic occupation coding, which can in turn affect exposure classification.[15] This is especially true for highly specific free-text occupation data that is used as input and for automatic coding methods at the six-digit SOC code level of classification compared to the two-digit SOC code level.[15] Several studies have found only modest agreement between manually and automatically coded occupations at the six-digit SOC code level.[26,29,30] Another study found that as the quality and standardization of input datasets increased, the concordance between manual and automatic coding results increased.[15] Unfortunately, occupational data for birth defects research often come from sources such as registries and birth certificates, which collect these data in nonstandardized free-text form. While machine learning has proven effective in many settings, caution is still needed regarding automatic occupational coding since the collection of occupational data is not standardized in birth defects research; thus, when resources are available, supplemental manual coding should continue to be implemented as the gold standard.

Our study has several strengths. Firstly, our study was conducted using data from the TBDR, one of the largest population-based birth defects registries in the United States. This large dataset allowed us to examine associations between several maternal occupations and risk of offspring clefts. Furthermore, while most previous studies assessing maternal occupation and offspring birth defect risk to date have been conducted using the NBDPS, our use of the TBDR data allowed us to examine this association in a novel study population. Additionally, our study population was limited to a well-defined group of cases with CLP/CP, allowing for ease of replication of our findings. Lastly, maternal occupations were defined using SOC codes, which allowed for unambiguous occupational coding categorization.

Our study should also be considered in light of some limitations. First, maternal occupation information was collected based on self-report, so there is potential for misclassification bias. However, since mothers are more likely to accurately report their occupations compared to other exposures (eg, alcohol use during pregnancy), the potential for this bias should be small.[2,31] Furthermore, maternal occupation data were missing for less than 5% of mothers in our dataset. An additional limitation was the exclusion of fetal deaths in our analyses since information on maternal occupation is not collected for Texas fetal death certificates. However, since only approximately 1% of our study population was excluded due to fetal deaths (data not shown), the likelihood of selection bias is very small. Although the strength of our study was the homogenous nature of our cases definition, a subsequent limitation is that our study only assessed one example birth defect. Additionally, due to the rarity of certain maternal occupations (eg, armed forces), certain analyses were not conducted for some maternal occupation groups. A larger, nationally representative sample size would allow for analyses among all major occupation groups and also improve generalizability of the study beyond Texas. Another important aspect to consider in interpreting our study's findings is the potential influence of unmeasured confounding factors. Since our study primarily focused on maternal occupation as a potential risk factor, we were unable to capture the entirety of influences that might contribute to the development of birth defects. Additionally, the birth certificate from which occupation data is obtained only asks for the mother's occupation and industry at the time the birth certificate is completed. There is no information regarding the length of time the mother has been employed at that occupation or information on the specific job-related exposures the mother may have encountered. Maternal exposure to risk factors during the first trimester of pregnancy has been found to be associated with orofacial clefts. However, we are unable to ascertain specific periods of occupational exposure during pregnancy due to the unavailability of such information in the birth certificate records. Future studies could benefit from including a broader range of variables to provide a more detailed understanding of the complex interplay between maternal occupation and birth defects.

In conclusion, our results support the notion that utilization of NIOCCS for automatic coding holds promise as a viable and feasible approach in conducting occupation-related analyses, especially considering the substantial advancements in its machine learning algorithms, resulting in an increase in automatic coding accuracy. Our study builds upon a relatively limited landscape of research investigating occupations and occupational exposures as potential risk factors for birth defects. Further advancements in automatic coding methodologies will likely improve occupational exposure ascertainment and its application in birth defects research.

ACKNOWLEDGMENTS

This work was supported by T42 OH008421/OH/NIOSH CDC HHS/United States training grant and funding from the Centers for Disease Control and Prevention (CDC), as part of a birth defects surveillance cooperative agreement with the Texas Department of State Health Services (TX DSHS) (NU50DD000102). The TX DSHS Birth Defects Epidemiology and Surveillance Branch is supported in part by Maternal and Child Health Services Title V Block Grant funds from the Health Resources and Services Administration (HRSA). The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement by, CDC, TX DSHS, or HRSA.

REFERENCES

1. Park R, Grant D, Jans M, Frausto M, Rauch J. Comparing manual and automated industry and occupation coding: accuracy and cost from the perspective of the California Health Interview Survey. 2015. In Proceedings of the Survey Research Methods Section, American Statistical Association. Retrieved from http://www.asasrms.org/Proceedings/y2015/files/234248.pdf. Accessed August 15, 2024.
2. Lin S, Herdt-Losavio ML, Chapman BR, et al. Maternal occupation and the risk of major birth defects: a follow-up analysis from the National Birth Defects Prevention Study. *Int J Hyg Environ Health* 2013;216: 317–323.
3. Desrosiers TA, Herring AH, Shapira SK, et al. Paternal occupation and birth defects: findings from the National Birth Defects Prevention Study. *Occup Environ Med* 2012;69:534–542.
4. Herdt-Losavio ML, Lin S, Chapman BR, et al. Maternal occupation and the risk of birth defects: an overview from the National Birth Defects Prevention Study. *Occup Environ Med* 2010;67:58–66.
5. Schnitzer PG, Olshan AF, Erickson JD. Paternal occupation and risk of birth defects in offspring. *Epidemiology* 1995;6:577–583.
6. Kim J, Langlois PH, Mitchell LE, Agopian AJ. Maternal occupation and the risk of neural tube defects in offspring. *Arch Environ Occup Health* 2018;73: 304–312.
7. Kim J, Langlois PH, Herdt-Losavio ML, Agopian AJ. A case-control study of maternal occupation and the risk of orofacial clefts. *J Occup Environ Med* 2016;58:833–839.
8. Siegel M, Rocheleau CM, Johnson CY, et al. Maternal occupational oil mist exposure and birth defects, National Birth Defects Prevention Study, 1997–2011. *Int J Environ Res Public Health* 2019;16:1560.
9. Cordier S, Bergeret A, Goujard J, et al. Congenital malformation and maternal occupational exposure to glycol ethers. Occupational exposure and congenital malformations working group. *Epidemiology* 1997;8:355–363.
10. Spinder N, Bergman JEH, Boezen HM, Vermeulen RCH, Kromhout H, de Walle HEK. Maternal occupational exposure and oral clefts in offspring. *Environ Health* 2017;16:83.
11. NIOSH. NIOSH Industry and Occupation Computerized Coding System (NIOCCS). U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health, Division of Field Studies & Engineering, Health Informatics Branch. https://csams.cdc.gov/nioccs/About.aspx. Accessed September 8, 2022.
12. Taylor JA, Frey LT. The need for industry and occupation standards in hospital discharge data. *J Occup Environ Med* 2013;55:495–499.
13. Patel MD, Rose KM, Owens CR, Bang H, Kaufman JS. Performance of automated and manual coding systems for occupational data: a case study of historical records. *Am J Ind Med* 2012;55:228–231.
14. Burstyn I, Slutsky A, Lee DG, Singer AB, An Y, Michael YL. Beyond crosswalks: reliability of exposure assessment following automated coding of free-text job descriptions for occupational epidemiology. *Ann Occup Hyg* 2014;58:482–492.
15. Roberts B, Shkembi A, Smith LM, Neitzel RL. Beware the Grizzlyman: a comparison of job- and industry-based noise exposure estimates using manual coding and the NIOSH NIOCCS machine learning algorithm. *J Occup Environ Hyg* 2022;19:437–447.
16. Miller E. Evaluation of the Texas Birth Defects Registry: an active surveillance system. *Birth Defects Res A Clin Mol Teratol* 2006;76:787–792.
17. Rasmussen SA, Olney RS, Holmes LB, et al. Guidelines for case classification for the National Birth Defects Prevention Study. *Birth Defects Res A Clin Mol Teratol* 2003;67:193–201.
18. Savitz DA, Whelan EA, Rowland AS, Kleckner RC. Maternal employment and reproductive risk factors. *Am J Epidemiol* 1990;132:933–945.
19. Joffe M. Biases in research on reproduction and women's work. *Int J Epidemiol* 1985;14:118–123.
20. 2010 SOC User Guide. https://www.bls.gov/soc/soc_2010_user_guide.pdf. Accessed September 15, 2022.
21. SAS Institute Inc. SAS/STAT® 9.4 User's Guide. Cary, NC: SAS Institute Inc.; 2013.
22. Siegel MR, Rocheleau CM, Broadwater K, et al. Maternal occupation as a nail technician or hairdresser during pregnancy and birth defects, National Birth Defects Prevention Study, 1997–2011. *Occup Environ Med* 2022;79: 17–23.
23. Brender J, Suarez L, Hendricks K, Baetz RA, Larsen R. Parental occupation and neural tube defect-affected pregnancies among Mexican Americans. *J Occup Environ Med* 2002;44:650–656.
24. Lorente C, Cordier S, Bergeret A, et al. Maternal occupational risk factors for oral clefts. Occupational exposure and congenital malformation working group. *Scand J Work Environ Health* 2000;26:137–145.
25. Occupational Outlook Handbook: Janitors and Building Cleaners. Washington, DC: U.S. Department of Labor. Retrieved from https://www.bls.gov/ooh/building-and-grounds-cleaning/janitors-and-building-cleaners.htm. Accessed October 30, 2022.

26. 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 Inform* 2016;4:e5.

27. Wuellner S, Levenson C. Occupation and industry data quality among select notifiable conditions in Washington State. *J Public Health Manag Pract* 2024; 30:36–45.

28. Dill J, Duffy M. Structural racism and black Women's employment in the US health care sector. *Health Aff (Millwood)* 2022;41:265–272.

29. Russ DE, Ho KY, Colt JS, et al. Computer-based coding of free-text job descriptions to efficiently identify occupations in epidemiological studies. *Occup Environ Med* 2016;73:417–424.

30. Buckner-Petty S, Dale AM, Evanoff BA. Efficiency of autocoding programs for converting job descriptors into standard occupational classification (SOC) codes. *Am J Ind Med* 2019;62:59–68.

31. Teschke K, Olshan AF, Daniels JL, et al. Occupational exposure assessment in case-control studies: opportunities for improvement. *Occup Environ Med* 2002;59:575–593; discussion 594.