Beware the Grizzlyman: A comparison of job- and industry-based noise exposure estimates using manual coding and the NIOSH NIOCCS machine learning algorithm

Benjamin Roberts\textsuperscript{a}, Abas Shkembi\textsuperscript{b}, Lauren M. Smith\textsuperscript{b}, and Richard L. Neitzel\textsuperscript{b}

\textsuperscript{a}Cardno ChemRisk, Chicago, Illinois; \textsuperscript{b}Department of Environmental Health Sciences, University of Michigan School of Public Health, Ann Arbor, Michigan

**ABSTRACT**

Recently, the National Institute for Occupational Safety and Health (NIOSH) released an updated version of the NIOSH Industry and Occupation Computerized Coding System (NIOCCS), which uses supervised machine learning to assign industry and occupational codes based on provided free-text information. However, no efforts have been made to externally verify the quality of assigned industry and job titles when the algorithm is provided with inputs of varying quality. This study sought to evaluate whether the NIOCCS algorithm was sufficiently robust with low-quality inputs and how variable quality could impact subsequent job estimated exposures in a large job-exposure matrix for noise (NoiseJEM). Using free-text industry and job descriptions from \( \textgreater \) 700,000 noise measurements in the NoiseJEM, three files were created and input into NIOCCS: (1) N1, “raw” industries and job titles; (2) N2, “refined” industries and “raw” job titles; and (3) N3, “refined” industries and job titles. Standardized industry and occupation codes were output by NIOCCS. Descriptive statistics of performance metrics (e.g., misclassification/discordance of occupation codes) were evaluated for each input relative to the original NoiseJEM dataset (N0). Across major Standardized Occupational Classifications (SOC), total discordance rates for N1, N2, and N3 compared to N0 were 53.6%, 42.3%, and 5.0%, respectively. The impact of discordance on the major SOC group varied and included both over- and under-estimates of average noise exposure compared to N0. N2 had the most accurate noise exposure estimates (i.e., smallest bias) across major SOC groups compared to N1 and N3. Further refinement of job titles in N3 showed little improvement. Some variation in classification efficacy was seen over time, particularly prior to 1985. Machine learning algorithms can systematically and consistently classify data but are highly dependent on the quality and amount of input data. The greatest benefit for an end-user may come from cleaning industry information before applying this method for job classification. Our results highlight the need for standardized classification methods that remain constant over time.

**KEYWORDS**

Big data; machine learning; noise; SOC

---

**Introduction**

Epidemiological studies can be difficult to undertake due to the time and costs associated with accurately assessing exposure over an extended period of time. In many cases, researchers only have limited access to research subjects and have difficulty assessing how an individual’s exposure changes over the duration of the study. Job-exposure matrices (JEMs) are an invaluable tool for researchers to quickly estimate exposures for workers without collecting primary data. These tools have been widely used to assess exposure in occupational epidemiology studies (Kauppinen et al. 1998; Koeman et al. 2013; Kauppinen et al. 2014).

Recently, a large JEM of occupational noise measurements was constructed using data provided by U.S. and Canadian governmental agencies, private employers, and the published literature. The large bulk of these measurements was made in accordance with the U.S. Occupational Safety and Health Administration (OSHA) noise standard for measuring 8-hr time-weighted averages (TWAs) (Roberts et al. 2018) and was prominently collected in the mining (Roberts...
et al. 2017) and manufacturing (Sayler et al. 2019) sectors, which traditionally have the highest rates of noise-induced hearing loss (NIHL) (Masterson et al. 2015). Measurements in this JEM were organized by industry using the 2012 North American Industry Classification System (NAICS), while job titles were organized using the 2010 Standard Occupational Classification (SOC) system (BLS 2021). All job and industry codes were manually assigned by two experienced industrial hygienists. A subsequent meta-analysis of the NoiseJEM found that 63% of SOCs had moderate to high heterogeneity, and 51% of SOCs had a pooled point estimate greater than 85 dBA (Cheng et al. 2018). The SOCs exhibiting moderate to high heterogeneity most often contained very few measurements. In an attempt to address this issue, parametric Bayesian imputation was used to take advantage of the hierarchical structure of the SOC system. Related job titles were used to help estimate the posterior mean and credible interval for each broad SOC group (Roberts et al. 2018). Of the 444 broad SOC groups, 85 (19%) had a posterior mean exposure greater than 85 dBA and 10 (2%) had a posterior mean exposure greater than 90 dBA (Roberts et al. 2018).

While considerable effort has been spent to verify the accuracy of the NoiseJEM and quantify uncertainty, it still suffer from the primary limitation that all JEMs suffer from, namely, misclassification of job titles. Because the NoiseJEM was constructed from noise measurements made by a variety of organizations over an extended period of time, there was significant variation in the job titles between the different sources of measurements. Recently, the National Institute for Occupational Safety and Health (NIOSH) released an updated version of the NIOSH Industry and Occupation Computerized Coding System (NIOCCS) that uses a deterministic supervised machine learning algorithm to assign NAICS and SOC codes based on the provided free-text information (NIOSH 2021). Further, NIOCCS system also classifies jobs and industry into broader industry and occupational codes for linking to the U.S. Census job and industry classifications (US Census Bureau 2021). The NIOCCS system allows for single entry or data file entry linkage, but it provides important crosswalks (in text and numeric form) to navigate between NAICS, SOCs, and Census definitions. While the new NIOCCS system gives researchers a convenient, low-cost, and significantly quicker tool to standardize job and industry titles, there has been no effort to externally evaluate the accuracy of assigned NAICS and SOC codes and how it varies by data input completeness/cleanliness.

To address the need for simple and consistent job and industry classifications, this study examined the level of disagreement between manually assigned job classifications and those assigned by the NIOCCS system. This study also examined how the level of “cleanliness” of the dataset facilitated the performance of the NIOCCS system. We assume that “cleaner” industry and job descriptions will improve the performance of the classification system over uncleaned free-text descriptions. In addition, this research examined whether or not manually assigned job and industry classifications would result in different job- and industry-specific exposure estimates when compared to the job and industry classifications assigned by the NIOCCS system. This was accomplished using three levels of data: (1) “raw,” (2) “partially refined,” and (3) “refined” to compare the NIOCCS system classification to the manually classified Noise JEM. Second, the impact that classification of each dataset by the NIOCCS system had on noise exposure estimates by industry and job codes was evaluated.

Methods

The underlying dataset of the large NoiseJEM used in this study is freely available online (https://noise.shinyapps.io/noiseJEM/). All data processing, integration, and analysis was performed in R v4.0.2 (Boston, MA). While the NoiseJEM contains measurements made in accordance with both OSHA and NIOSH, only measurements made using the OSHA Permissible Exposure Limit (PEL) standard, which make up the majority of the data set, were used for this analysis. Each of the measurements compiled into the NoiseJEM were collected from a specific job title and, therefore, at minimum include a free-text job title/description and sometimes include a free-text industry description. However, none of the data collected to compile the NoiseJEM was associated with standardized job or industry classification codes. Over 1,000 hr went into manually classifying the job title description for each noise measurement by the 2010 SOC (standardized occupational classification) structure. This analysis does not include legal occupations (major SOC group 23-0000) or military specific occupations (55-0000), which are not present in the NoiseJEM, leaving 748,586 noise measurements for analysis.

NIOCCS inputs

Three different combinations of four free-text descriptions of industries and job titles from the NoiseJEM
were input into the NIOCCS auto-coding system: one each of the original industry and job title descriptions provided during the JEM construction and one each of the manually classified industry and job title descriptions. Industry descriptions were manually classified using the 2012 NAICS structure and job title descriptions were manually classified using the 2010 SOC structure. These four variables were combined in three ways to compare the performance of the NIOCCS system: (1) the original industry and job title descriptions; (2) the manually classified industry descriptions and the original job title descriptions; and (3) the manually classified industry and job title descriptions. This allowed us to compare the performance of an entirely “raw” input, a “partially refined” input, and “refined” input, respectively, to each other.

The underlying dataset contained some noise measures in which no industry was provided in the original data. Overall, this accounted for 481,648 individual noise measures in the NoiseJEM (~48%). These entries with missing industry descriptions were maintained for input into the NIOCCS system to evaluate how the NIOCCS algorithm performed on missing data. The “raw” and “partially refined” inputs contained 55,998 and 68,908 unique combinations of industry and job title descriptions, respectively. There were more unique combinations of industry and job title descriptions in the “partially refined” than in the “raw” because many of the rows with missing industry descriptions were filled in during manual classification using information in the job title descriptions. The “refined” input contained substantially less, with 7,760 unique combinations. Each input dataset was then converted into a tab-separated .txt file and uploaded into the online NIOCCS system. Once the auto-coding was done for each input dataset, the auto-coded files were merged back together with the data from the NoiseJEM using a unique identifier.

**NIOCCS outputs**

In total, three alternative datasets for comparison with the reference dataset (the Noise JEM, denoted as N0). These were: N1 (“raw”), using the original industry (“raw”) and job descriptions; N2 (“partially refined”), using the manually recoded descriptor of four-digit NAICS code and original job descriptions; and N3 (“refined”), using the manually recoded four-digit NAICS code descriptor and the manually recoded broad-SOC code descriptor. Figure 1 provides an example data from the NoiseJEM that was input into the NIOCCS system in three ways depending on the dataset (i.e., N1, N2, or N3). The NIOCCS system reports the following metrics: mean probability that the NAICS or SOC code was correctly assigned, respectively, and the percentage of NAICS or SOC
codes that could not be assigned, respectively. In addition, the percent total discordance between the SOC codes assigned by the NIOCCS system and the manually assigned SOC codes was calculated for major, minor, broad, and detailed SOC levels.

**Statistical analysis**

To facilitate the comparison, NAICS and SOC classifications in the reference dataset, N0, are considered “correct” and will be referred to as such. Disagreement (rather than agreement) was measured by calculating individual discordance for each major SOC group within every alternative dataset (N1, N2, and N3) relative to the reference dataset. Discordance rates were used because we sought to identify cases where the algorithm diverged from the manually assigned classifications. Reliability of each alternative dataset was measured using Cohen’s Kappa ($\kappa$) to account for agreement simply due to chance.

To assess the bias and precision of each alternative dataset on noise estimates, average noise levels for each major SOC group by dataset were calculated along with mean errors and median squared errors for comparison to the reference dataset. A sensitivity analysis was performed for three major SOC groups at the broad SOC level, although rather than using squared errors for analysis, differences relative to the reference dataset were computed and analyzed for each broad SOC group within the alternative datasets. Median and IQRs were used to summarize metrics (e.g., discordance) skewed to the right.

**Results**

**NIOCCS performance**

In total, 990,433 measurements (excluding major SOC group 55-0000) were included in this analysis. The NIOCCS algorithm reported that dataset N2 had the highest mean probability of generating a correct NAICS code (89%), while N3 had the highest mean probability of generating a correct SOC code (84.5%) (Table 1). Similarly, dataset N2 had the lowest percent of unknown NAICS codes (0.1%) and N3 had the lowest percent of unknown SOC codes (2.3%). When compared to the reference dataset (N0), the percentage of discordance decreased as better information was supplied to the algorithm (i.e., from N1, to N2, and to N3). The only exception to this trend occurred at the detailed SOC level where dataset N1 had a slightly lower rate of total discordance compared to dataset N2. However, as would be expected, the least granular level of SOC (major), which has the fewest number of groups, had the lowest level of total discordance compared to the more granular levels of the SOC hierarchy. The results of the sensitivity analysis (examining only measurements with full information for industry) indicated that the lack of industry description did not have any meaningful difference in discordance rates. As would be expected measures of reliability followed a similar trend to discordance, with the N3 dataset having a substantially higher Major, Minor, and Broad SOC reliability ($\kappa = 0.68-0.91$) compared to the N1 and N2 datasets. The N2 dataset had marginally higher reliability compared to the N1 dataset at each of these SOC levels. Reliability was comparably low for each alternative dataset at the detailed SOC level ($\kappa = 0.16, 0.13, 0.25$ for N1, N2, and N3 dataset, respectively).

**Major SOC discordance**

Rates of discordance between each of the alternative datasets and the reference dataset were calculated for each major SOC group and are shown in Figure 2. Median discordance (as a proportion) among major SOC groups was substantially lower in the N3 dataset (median = 0.08; IQR = 0.18) compared to the N1 (0.64; 0.37) and N2 (0.68; 0.36). While the N3 dataset outperformed the N1 and N2 datasets, the rates of discordance did not maintain a uniform distribution by major SOC code, indicating bias in performance. Specifically, even in the “refined” N3 dataset, 78.9% of 104 inputs from the major SOC group “31-0000” (Healthcare Support Occupations) were classified into

---

**Table 1. Average performance metrics of the NIOCCS system by input dataset.**

| Metric                  | N1     | N2     | N3     |
|-------------------------|--------|--------|--------|
| NIOCCS output           |        |        |        |
| Mean NAICS probability  | 0.79   | 0.89   | 0.85   |
| % NAICS unknown         | 55.9   | 0.1    | 0.3    |
| Mean SOC probability    | 0.80   | 0.78   | 0.84   |
| % SOC unknown           | 31.8   | 9.4    | 2.3    |
| Measure of reliability  |        |        |        |
| Total major SOC discordance (%) | 53.6  | 42.3   | 5.0    |
| Without missing industry | 54.8  | 45.9   | 3.3    |
| Total minor SOC discordance (%) | 75.4  | 52.7   | 10.2   |
| Without missing industry | 78.0  | 47.4   | 6.9    |
| Total broad SOC discordance (%) | 77.6  | 61.2   | 23.2   |
| Without missing industry | 80.8  | 60.3   | 19.9   |
| Total detailed SOC discordance (%) | 83.1  | 86.3   | 74.5   |
| Without missing industry | 84.6  | 95.3   | 81.3   |
| Measure of reliability (\kappa) | 0.29  | 0.34   | 0.91   |
| Total major SOC reliability | 0.19  | 0.29   | 0.84   |
| Total minor SOC reliability | 0.19  | 0.27   | 0.68   |
| Total broad SOC reliability | 0.16  | 0.13   | 0.25   |

*Note the high % of NAICS unknown for N1 is due to high occurrence of completely missing industry information in that dataset.*
the wrong major SOC group, substantially higher than the overall major SOC code discordance of the N3 dataset (5%). Supplementary material Table S4 summarizes, for each major SOC group, which major SOC group the majority of misclassifications are classified into by the NIOCCS algorithm. Major SOC groups 39-0000 (Personal Care and Service Occupations), 27-0000 (Arts, Design, Entertainment, Sports, and Media Occupations), 15-0000 (Computer and Mathematical Occupations), 49-0000 (Installation, Maintenance, and Repair), and 13-0000 (Business and Financial Operations Occupations) also had substantially higher rates of discordance (>20%) than the overall rate of discordance.

A similar pattern of non-uniform discordance was observed in the N1 and N2 datasets, with some major SOC codes having substantially higher discordance rates and some having substantially lower relative to the respective overall discordance rate (Figure 2). The discrepancy between each individual major SOC grouping discordance rate and the total major SOC discordance rate among each alternative dataset varied. The median discrepancy (i.e., the difference between each major SOC group rate of discordance and the total major SOC discordance for each alternative dataset) was lowest among the N3 dataset (3%), with the N1 and N2 datasets having greater discrepancy (10% and 25%, respectively). Tables S1–S3 in the supplementary material provide a detailed summary of the discordance between datasets at the major, minor, and broad SOC level.

While a detailed analysis of the topic is beyond the scope of this paper, it is nevertheless important to acknowledge the potential impact that changes over time in classification methods and schema, or in reporting mechanisms and accuracy, can have on temporally distributed data such as those incorporated into the NoiseJEM. Supplementary material Figure S1 shows how the discordance rates for each major SOC group vary by year between the reference and alternative datasets. Generally, discordance rates across the three datasets were stable over time, with the exception of the years prior to 1985, which showed greater variability.
**Table 2.** Average noise levels for each major SOC group.

| Major SOC code | Job title | N     | Mean ± SD | N     | Mean ± SD | N     | Mean ± SD | N     | Mean ± SD |
|----------------|-----------|-------|-----------|-------|-----------|-------|-----------|-------|-----------|
| 11-0000        | Management| 1,380 | 86.1 ± 8.8| 4,102 | 80.6 ± 8.4| 1,453 | 81 ± 10.3 | 1,620 | 85.5 ± 9.1 |
| 13-0000        | Business and Financial Operations | 39 | 78.8 ± 10.7| 615 | 79.8 ± 10 | 524 | 80.7 ± 10.2 | 48 | 79.1 ± 9.7 |
| 15-0000        | Computer and Mathematical | 25 | 78.9 ± 7.3 | 120 | 85.4 ± 9.5 | 38 | 78.5 ± 9.5 | 24 | 74.9 ± 8.8 |
| 17-0000        | Architecture and Engineering | 7,176 | 80.1 ± 7.7 | 1,457 | 81.1 ± 8.4 | 2,747 | 82 ± 7.3 | 7,161 | 80.1 ± 7.7 |
| 19-0000        | Life, Physical, and Social Science | 776 | 79.6 ± 8.0 | 2,454 | 78.4 ± 8.7 | 8,816 | 78.9 ± 7.7 | 775 | 79.9 ± 8.1 |
| 21-0000        | Community and Social Service | 7 | 79.8 ± 11.6 | 418 | 81.7 ± 11.3 | 28 | 81.9 ± 8 | 26 | 79.2 ± 8.5 |
| 25-0000        | Education, Training, and Library | 165 | 82.8 ± 9.2 | 229 | 83.9 ± 8 | 249 | 82.5 ± 8.6 | 164 | 83.2 ± 8.7 |
| 27-0000        | Arts, Design, Entertainment, Sports, and Media | 238 | 85.1 ± 9.2 | 563 | 87.3 ± 8.5 | 148 | 85.5 ± 9 | 193 | 85.8 ± 8.8 |
| 29-0000        | Healthcare Practitioners and Technical | 220 | 78.5 ± 11.5 | 743 | 85.4 ± 10.2 | 167 | 79.2 ± 10.1 | 264 | 78.9 ± 11 |
| 31-0000        | Healthcare Support | 67 | 82.2 ± 10 | 78 | 84.1 ± 9.6 | 38 | 83.2 ± 8.2 | 26 | 75.1 ± 6.5 |
| 33-0000        | Protective Service | 480 | 83.3 ± 10.2 | 863 | 87.6 ± 9.2 | 366 | 84.8 ± 10.4 | 469 | 83.6 ± 10.2 |
| 35-0000        | Food Preparation and Serving Related | 319 | 82.9 ± 8.5 | 1,309 | 87.7 ± 8.2 | 332 | 83.6 ± 8.3 | 274 | 82.8 ± 8.2 |
| 37-0000        | Building and Grounds Cleaning and Maintenance | 1,675 | 87.1 ± 7.9 | 4,414 | 83.6 ± 9.2 | 1,473 | 85.4 ± 9.6 | 2,111 | 85.9 ± 8.4 |
| 39-0000        | Personal Care and Service | 178 | 86.9 ± 9.6 | 1,116 | 88 ± 9.2 | 234 | 85.3 ± 9.6 | 101 | 84.6 ± 10.4 |
| 41-0000        | Sales and Related | 935 | 84.8 ± 9 | 459 | 88.3 ± 9.4 | 676 | 84.2 ± 9.4 | 967 | 85.9 ± 8.9 |
| 43-0000        | Office and Administrative Support | 2,038 | 78.3 ± 9 | 12,701 | 80.7 ± 8.5 | 5869 | 81.7 ± 8.3 | 2090 | 78.9 ± 9 |
| 45-0000        | Farming, Fishing, and Forestry | 1,384 | 90.4 ± 8.6 | 2,015 | 84.5 ± 9 | 2149 | 89.5 ± 8.7 | 9764 | 92.9 ± 7.3 |
| 47-0000        | Construction and Extraction | 469,227 | 82.8 ± 9.2 | 211,792 | 83.6 ± 9.1 | 369,630 | 83.3 ± 9 | 486,098 | 82.8 ± 9.2 |
| 49-0000        | Installation, Maintenance, and Repair | 44,769 | 81.3 ± 8.9 | 28530 | 80.7 ± 8.3 | 21,476 | 81.1 ± 8.6 | 35,299 | 81.6 ± 8.8 |
| 51-0000        | Production | 135,533 | 86.3 ± 8.6 | 178,602 | 84.7 ± 9.1 | 125,826 | 86.8 ± 8.7 | 123,254 | 85.9 ± 8.4 |
| 53-0000        | Transportation and Material Moving | 81,955 | 82.6 ± 8.7 | 227,827 | 81.9 ± 9.0 | 186,959 | 81.4 ± 9.1 | 76,867 | 82.5 ± 8.7 |

**Effect on exposure estimates at the major SOC level**

After filtering for only OSHA PEL measurements, 748,586 measurements were included in this portion of analysis. Average noise levels for major SOC groups in each of the datasets are presented in Table 2. The difference between these average noise levels in alternative datasets compared to the reference dataset were calculated and are shown in Supplementary material Figure S2. The “raw” dataset, N1, has the most error relative to mean estimates in the reference dataset, and the “partially refined” dataset, N2, substantially reduces the amount of relative error. The squared errors of the major SOC groups in N2 dataset were substantially lower relative to the reference dataset (median = 0.87; IQR = 2.38) than were those of the N1 dataset (3.59; 15.4). This demonstrates that there was much less bias in average major SOC group estimates in the “partially refined” dataset compared to the “raw” dataset. The “refined” dataset outperformed the other two datasets with regard to squared error (0.13; 0.38). Regardless, it is important to note a handful of major SOC groups in the N3 dataset which had relatively large error compared to the reference dataset, such as 31-0000 (Healthcare Support Occupations), 15-0000 (Computer and Mathematical Occupations), and 39-0000 (Personal Care and Service Occupations). Overall, noise levels for the N1 and N2 datasets were, on average, negatively biased (mean difference of –0.93 dBA and –0.05 dBA, respectively), while the N3 dataset was positively biased (mean difference of 0.52 dBA) as compared to the N0 dataset.

Three major SOC groupings were selected to examine the impact on average OSHA PELs annually by quantity of data: (1) “47-0000” (Construction and Extraction Occupations), which accounted for around 63.3% (469,227) of measurements in the NoiseJEM; (2) “45-0000” (Farming, Fishing, and Forestry Occupations), which accounted for around 0.1% (1,384) of measurements in the NoiseJEM; and (3) “39-0000” (Personal Care and Service Occupations), which contained a mere 178 (0.02%) measurements in the NoiseJEM. Annual OSHA PELs were calculated for each of the three alternative datasets and the reference dataset and are presented in Figure 3. This figure demonstrates that for an occupation group like “47-0000” which contains hundreds of thousands of measurements, the overall annual trend of noise exposure remains relatively consistent among the four datasets, indicating the annual trend in this group has a high level of robustness to misclassification from the NIOCCS algorithm. The median variability between the four datasets within a year in the “47-0000” group was minimal, 0.61 (IQR = 0.48) dBA. In contrast, when examining temporal trends of the “45-0000” occupation group, the average yearly variability between the four datasets was substantially higher at 3.27 (2.78) dBA compared to that of the “47-0000” group. Meanwhile, groups with fewer measurements like “39-0000,” which contained around 150
measurements, were the least robust to misclassification by the NIOCCS algorithm and had the greatest discordance between the datasets. This is further highlighted by the median yearly variability between the four datasets (4.29; 3.33 dBA), which was even higher than the median yearly variability of “45-0000.”

**Effect on exposure estimates at the broad SOC level**

Annual OSHA PELs were calculated for broad SOC groups within the three major SOC groups selected above for each of the three alternative datasets and the reference dataset. The difference between the reference and each of the alternative datasets was calculated and presented in Figure 4. A full breakdown of each mean difference at the broad SOC level is presented in Supplementary material Table S5. Recreation and Fitness Workers (broad SOC group “39-9030”) were observed to have the highest mean difference relative to the N0 dataset when compared to the N2 and N1 datasets (–26.6 dBA and –23.1 dBA, respectively). Mean differences at the broad SOC level ranged from –26.6 to 11.6 dBA among major SOC group “39-0000,” –10.9 dBA to 7.1 dBA for “45-0000,” and

---

**Figure 3.** Annual noise levels (OSHA PELs) of the reference dataset (N0) and the alternative datasets (N1 – raw dataset, N2 – partially refined dataset, N3 – refined dataset) for major SOC groups (from top to bottom) 47-0000 (“Construction and Extraction Occupations”), 45-0000 (“Farming, Fishing, and Forestry”), and 39-0000 (“Personal Care and Service Occupations”). Left-sided panel displays traditional line plots; right-sided panel displays smoothed line plots. Black, solid line indicates the “true” average measurement (N0).
—9.6 to 9.1 dBA for “47-0000.” These ranges were substantially higher than those observed at the major SOC level (Supplementary material, Figure S2). Overall, the N3 dataset was positively biased for broad SOC groups among the “47-0000” and “39-0000” major SOC groups (mean difference of 0.5 dBA and 2.5 dBA, respectively) and negatively biased for the “45-0000” major SOC group. The N2 dataset was a bit more even among each of the major SOC groups, but was still slightly negatively biased with an overall average of —1.75 dBA. For the N1 dataset, the broad SOC groups among the “39-0000” were the mostly negatively biased, with a mean difference of —1.2 dBA.

**Discussion**

This analysis demonstrates the challenge of acquiring occupational exposure data from a variety of sources and synthesizing that data into a uniform dataset that can be used for analysis. Machine learning represents one tool that can assist humans with classifying and grouping data in a way that would likely be impossible with manual efforts (He et al. 2021). While numerous examples of machine learning “correctly” classifying data exist, the results of this analysis suggest great care should be taken when utilizing algorithms for grouping occupational exposure data. As demonstrated here, the quality of the input data can make a significant impact on the assignment of occupational or industry groups, which in turn can impact the resulting distribution of exposures across a group.

Dataset N1 represented the “worst-case” example of data where many job or industry titles were missing, incomplete, or otherwise of low quality. Presumably, any data submitted to the NIOCCS system would undergo some level of data validation; however, this...
dataset represented the job and industry information as it was received during the compilation of the NoiseJEM. As would be expected, this dataset had the highest discordance with the manually assigned SOC and NAICS codes. The discordance was higher in major SOCs that contained less data. On the major SOC level, this resulted in major SOC means differing by as much as ±5 dB. On the broad SOC level these differences ranged between −27 and 12 dBA. Differences of this magnitude can have huge implications in practice and research, as OSHA’s noise standard, which was used to collect the measurements used in this analysis, considers a 5 dB increase to represent a doubling of exposure which requires that the allowable exposure duration be cut in half. This is in contrast to most other noise standards that consider a 3 dB increase to represent a doubling of exposure.

The difference between mean noise exposures at the major SOC level in the NoiseJEM decreased when compared to dataset N2 and even further when compared to dataset N3. However, there were some notable exceptions to this trend, as the difference in estimated noise levels for major SOC “31-0000” (“Healthcare Support Occupations”) was greatest between the NoiseJEM and dataset N3. This indicates that even relatively well-curated data supervised learning algorithms such as the NIOCCS system can still mis- or differentially classify data. These misclassifications can potentially result in drastically different exposure estimates even at the least granular level of grouping. As would be expected, however, large differences in exposure estimates were pervasive at more granular levels of grouping. A notable example is broad SOC “39-9030” (“Recreation and Fitness Workers”) where the difference in estimated average noise level in the N2 dataset was 26.6 dBA lower than in the reference dataset, N0. Overall, while the “partially refined” N2 and “refined” N3 datasets substantially improved on the differences in estimates in the “raw” N1 dataset. At times, the N2 dataset even outperformed the N3 dataset. This suggests that there may be a large benefit to refining industry inputs into the NIOCCS system, even if the job titles are not refined. The benefit of refining the job titles may not be as necessary as that of industries.

Burstyn et al. (2014) presented the results of a Java-based application that converted free-text information into 2010 SOC codes. The authors used three different datasets to compare exposure metrics for a dataset that was manually coded (considered the gold standard) and those coded by the Java application. The correlation between exposure metrics in the two datasets of physical demand scores was moderate ($r = 0.5$), while the agreement for polycyclic aromatic hydrocarbons (PAHs) was lower ($\kappa = 0.29$). The authors noted that the use of a purely automated coding system be “disastrous” if it were the sole classifier of exposure, and suggested that manual coding be used to improve automated coding systems, as we have done here. An example of using manual coding in a structured system was described by De Matteis et al. (2017), who developed the occupations self-coding automatic recording (OSCAR) system that guided individuals through a decision tree that was linked to SOC codes. Moderate agreement between the four-digit ($\kappa = 0.45$) and one-digit ($\kappa = 0.64$) SOC codes assigned by OSCAR and a manual coder. Similarly, this study found the highest Kappa scores for the broadest SOC categories with the Kappa scores decreasing as the detail of the SOCs increased. These results from Burstyn et al. (2014) and De Matteis et al. (2017) support the findings of this study that found improved free-text quality and less specific job titles improve the agreement between manually coded and automatically coded job titles.

Other studies have reached similar conclusions on the utility of auto-coding systems. For example, Russ et al. (2016) found agreement between coding methods to be 44.5% at the six-digit SOC level and 76.3% at the two-digit level. Similarly, Buckner-Petty et al. (2019) noted that agreement between manually and automatically assigned SOC codes was modest at the six-digit level and stronger at the two-digit level. Schmitz and Forst (2016) evaluated an older version of the NIOCCS system and reported that the highest agreement between manually and automatically coded SOC codes was at the two-digit level and decreased as the SOC codes became more specific. While Kappa and correlation statistics were not calculated in this study, we found a similar trend where a better prepared dataset decreased the discordance between manually and auto-coded job titles (Supplementary material, Table S1) and as the job details become more detailed the level of discordance tends to increase (Supplementary material, Table S2).

The SOCs used in the NoiseJEM and NIOCCS were developed for tracking broad economic factors and were not intended to be used for occupational exposure assessments (BLS 2010). The issue of collecting informative and standardize job title information was identified by Lippmann et al. (1996) decades before terms like “machine learning” entered the occupational health lexicon, and in 2018, the National Academies of Science (NAS) highlighted the
need for an occupational health surveillance system to capture data in a standardized manner across state and federal agencies (NAS 2018). Unlike other fields, the amount of occupational exposure data is limited, which necessitates the use of supervised learning algorithms, that are in turn limited by the availability of high-quality data with which to train the algorithm. Our results do not suggest that machine learning algorithms such as the NIOCCS system have no place in classifying job and industry titles for occupational exposure data. Rather, our study emphasizes the need for occupational health professionals to adopt a harmonized system for classifying job titles, and other information, while conducting exposure assessments. Going forward, the occupational health community needs to make a concerted effort to ensure that occupational exposure data is stored in a consistent manner that will allow data to be used for a longer period of time.

Acknowledgment

All authors have disclosed all actual and potential conflicts regarding this manuscript and the nature of these conflicts. One of the authors (BR) is employed by Cardno ChemRisk, a consulting firm that provides scientific advice to the government, corporations, law firms, and various scientific/professional organizations. Cardno ChemRisk was not involved in this research and did not review this manuscript prior to submission. The remaining authors are employed by or are students at the University of Michigan and received no support, financial or otherwise, from Cardno ChemRisk.

Data availability

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

References

BLS. 2010. 2010 SOC user guide. Washington (DC): US Bureau of Labor Statistics.

BLS. 2021. Standard Occupational Classification. US Bureau of Labor Statistics; [accessed Oct 4]. https://www.bls.gov/soc/2010/home.htm.

Buckner-Petty S, Dale AM, Evanoff BA. 2019. Efficiency of autocoding programs for converting job descriptors into standard occupational classification (SOC) codes. Am J Ind Med. 62(1):59–68. doi:10.1002/ajim.22928

Burstyn I, Slutsky A, Lee DG, Singer AB, An Y, Michael YL. 2014. Beyond crosswalks: reliability of exposure assessment following automated coding of free-text job descriptions for occupational epidemiology. Ann Occup Hyg. 58(4):482–492. doi:10.1093/annhyg/meu006

Cheng W, Roberts B, Mukherjee B, Neitzel RL. 2018. Meta-analysis of job-exposure matrix data from multiple sources. J Expo Sci Environ Epidemiol. 28(3):259–274. doi:10.1038/es.2017.19

De Matteis S, Jarvis D, Young H, Young A, Allen N, Potts J, Darnton A, Rushton L, Cullinan P. 2017. Occupational self-coding and automatic recording (OSCAR): a novel web-based tool to collect and code lifetime job histories in large population-based studies. Scand J Work Environ Health. 43(2):181–186. doi:10.5271/siwh.e3613

He X, Zhao K, Chu X. 2021. AutoML: a survey of the state-of-the-art. Knowl-Based Syst. 212:106622. doi:10.1016/j.knosys.2020.106622.

Kauppinen T, Toikkanen J, Pukkala E. 1998. From cross-tabulations to multipurpose exposure information systems: a new job-exposure matrix. Am J Ind Med. 33(4):409–417. doi:10.1002/(SICI)1097-0274(199804)33:4<409::AID-AJIM12>3.0.CO;2-2

Kauppinen T, Uuksulainen S, Saalo A, Makinen I, Pukkala E. 2014. Use of the Finnish Information System on Occupational Exposure (FINJEM) in epidemiologic, surveillance, and other applications. Ann Occup Hyg. 58(3):380–396. doi:10.1093/annhyg/met074

Koeman T, Offermans NS, Christopher-de Vries Y, Slottje P, Van Den Brandt PA, Goldbohm RA, Kromhout H, Vermeulen R. 2013. JEMs and incompatible occupational coding systems: effect of manual and automatic recoding of job codes on exposure assignment. Ann Occup Hyg. 57(1):107–114. doi:10.1093/annhyg/mes046

Lippmann M, Gomez MR, Raws GM. 1996. Special report: data elements for occupational exposure databases: guidelines and recommendations for airborne hazards and noise. Appl Occup Environ Hyg. 11:1294–1311.

Masterson EA, Deddens JA, Themann CL, Bertke S, Calvert GM. 2015. Trends in worker hearing loss by industry sector, 1981-2010. Am J Ind Med. 58(4):392–401.

NAS. 2018. A smarter national surveillance system for occupational safety and health in the 21st century. Washington (DC): National Academies of Sciences.

NIOSH. 2021. 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; [accessed Oct 4]. https://www.cdc.gov/nioccs3/FAQ.aspx.

Roberts B, Cheng W, Mukherjee B, Neitzel RL. 2018. Imputation of missing values in a large job exposure matrix using hierarchical information. J Expo Sci Environ Epidemiol. 28(6):615–648. doi:10.1038/s41370-018-0037-x

Roberts B, Sun K, Neitzel RL. 2017. What can 35 years and over 700,000 measurements tell us about noise exposure in the mining industry? Int J Audiol. 56(sup1):4–12.

Russ DE, Ho KY, Colt JS, Armenti KR, Baris D, Chow WH, Davis F, Johnson A, Purdue MP, Karagas MR, et al. 2016. Computer-based coding of free-text job descriptions to efficiently identify occupations in epidemiological studies. Occup Environ Med. 73(6):417–424. doi:10.1136/oemed-2015-103152

Sayler SK, Roberts BJ, Manning MA, Sun K, Neitzel RL. 2019. Patterns and trends in OSHA occupational noise exposure
measurements from 1979 to 2013. Occup Environ Med. 76(2):118–124. doi:10.1136/oemed-2018-105041
Schmitz M, Forst L. 2016. 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. 4(1):e5. doi:10.2196/medinform.4839
US Census Bureau. 2021. Industry and occupation code lists & crosswalks; [accessed Oct 4]. https://www.census.gov/topics/employment/industry-occupation/guidance/code-lists.html.