Published in final edited form as:
J Expo Sci Environ Epidemiol. 2018 November ; 28(6): 615–648. doi:10.1038/s41370-018-0037-x.

Imputation of Missing Values in a Large Job Exposure Matrix using Hierarchical Information

Benjamin Roberts, PhD[A], **Wenting Cheng, PhD**[B], **Bhramar Mukherjee, PhD**[B], **Richard L Neitzel, PhD**[A]

[A]Department of Environmental Health Sciences, University of Michigan, School of Public Health, Ann Arbor, Michigan, USA

[B]Department of Biostatistics, University of Michigan, School of Public Health, Ann Arbor, Michigan, USA

Abstract

Job exposure matrices (JEMs) represent a useful and efficient approach for estimating occupational exposures. This study uses a large dataset of full-shift measurements and employs imputation strategies to develop noise exposure estimates for almost all broad level standard occupational classification (SOC) groups in the US. The JEM was constructed using 753 702 measurements from the government, private industry, and the published literature. Parametric Bayes imputation was used to take advantage of the hierarchical structure of the SOCs and the mean occupational noise exposures were estimated for all broad level SOCs, except those in major group 23–0000, for which no data were available. The estimated posterior mean for all broad SOCs was found to be 82.1 dBA with within- and between-major SOC variabilities of 22.1 and 13.8, respectively. Of the 443 broad SOCs, 85 were found to have an estimated mean exposure >85 dBA while 10 were >90 dBA. By taking advantage of the size and structure of the dataset we were able to employ imputation techniques to estimate mean levels of noise exposure for nearly all SOCs in the US. Possible sources of errors in the estimates include misclassification of job titles due to limited data, temporal variations that were not accounted for, and variation in exposures within the same SOC. Our efforts have resulted in an almost completely-populated noise JEM that provides a valuable tool for the assessment of occupational exposures to noise. Imputation techniques can lead to maximal use of available information that may be incomplete.

Keywords

personal exposures; epidemiology; empirical/statistical models; exposure modeling

Corresponding author: Richard Neitzel, PhD, University of Michigan, Department of Environmental Health Sciences, 1415 Washington Heights, 6611 SPH I, Ann Arbor MI 48109, 734-763-2870, rneitzel@umich.edu.

Conflict of Interest
The authors have no conflicts of interest to declare.

Supplementary Information
Supplementary information is available at JESEE’s website.

Background

Noise induced hearing loss (NIHL) is the most common workplace injury, affecting an estimated 11.4% of workers in the United States.[1] While it is difficult to quantify the economic costs of NIHL, the US Veterans Administration reported direct costs of $1.2 billion in 2006 on hearing disability and tinnitus in addition to $288 million spent annually by the Veterans Administration on hearing aids.[2,3] More recently, we have estimated the direct and indirect costs of preventable NIHL to be between $58 and $152 billion annually in the US, with a central estimate of $123 billion per year.[4] Thus it is reasonable to assume that NIHL has a substantial and underappreciated ongoing impact on the US economy. Despite the clear relationship between hazardous noise exposure (>85 dBA) and hearing loss it is estimated that more than 22 million US workers are exposed to hazardous levels of noise at work.[5,6]

While it is well-established that hazardous noise exposure causes NIHL, conducting occupational epidemiological studies to further elucidate and quantify this relationship is challenging. Ideally, prospective cohort studies would be implemented to follow workers and monitor their noise exposure for a decade or more until the onset of significant NIHL. However, the costs and time required to conduct a longitudinal study make this approach difficult and rare. Typically, researchers instead rely on retrospective cohort studies to assess the relationship between an occupational exposure and a disease.[7] In these retrospective studies it can be difficult to develop to accurately estimate exposures.[8] To overcome these difficulties researchers have increasingly relied on job exposure matrices (JEMs) to retrospectively assess occupational exposures.[7,9–13]

In its most basic form a JEM consists of two axes: one axis contains a list of jobs or job descriptions, and the other contains qualitative or quantitative information about the magnitude and/or prevalence an exposure.[7] A JEM can be further refined by adding further information on specific job tasks, and the time period of exposure. The main advantage of a JEM is that it allows the use of previously collected industrial hygiene measurement records that greatly simplify epidemiological exposure assessment. A well-constructed JEM also makes it possible to identify occupations and industries that have potentially high levels of an exposure so that additional assessment and targeted controls can be implemented to reduce potential exposures.

There are many issues that arise when using a JEM as an exposure assessment tool. The first is that exposure varies depending on both a worker’s job title and the industry that the worker is employed in.[14] Workers with similar job titles can have large differences in their exposures depending on the industry they are employed in. It has also been shown that the majority of purportedly homogeneously exposed groups (HEGs) of workers – often based on job title – in the same workplace had more than a 2-fold difference in exposures.[15] The second issue is that exposure typically vary over time for a worker in the same job as changes in their workplace lead to a change in exposure patterns.[7,15] Finally, data scarcity often necessitates the use of qualitative exposure measures, which reduce the statistical power of a JEM to detect an exposure-response relationship.[16]

The JEM we describe here consists of 753 702 full-shift occupational noise measurements made according to the Occupational Safety and Health Administration's (OSHA) Permissible Exposure Limit (PEL) for noise.[17] Our previous meta-analysis of a subset of 715 867 measurements included in this JEM found that 26.4% of 235 job titles had no heterogeneity across sources (literature, government and industry reported sources), while 63.0% of job titles were found to have moderate to high levels of heterogeneity[18]. Despite the size and scope of this dataset, many job titles still lack exposure information. The goal of this present study is to take advantage of the hierarchical structure of the job title system used in this JEM in order to develop imputation strategies to calculate estimates of exposure and variability for job titles in which no exposure information is available and then determine which job titles have an estimated exposure greater than the current OSHA action level (AL) of 85 dBA and PEL of 90 dBA.

Methodology

The JEM was constructed using OSHA[17] and Mine Safety and Health Administration (MSHA)[19] PEL measurements (i.e. a 90 dBA criterion level and threshold, and 5 dB time-intensity exchange rate) from government databases maintained by OSHA and MSHA, measurements from the published literature, and measurements submitted by private industry (Table 1). Details about the data cleaning process for the JEM have been described elsewhere.[18,20] Briefly, data was received from the various sources in an electronic format, typically a Microsoft Excel file (Redmond, WA). The data was imported in to STATA 14 (College Station, TX) for data cleaning. Industry information was first coded using the 2012 North American Industrial Classification System (NAICS) from the US Census Bureau.[21] Using information on the industry of employment and job titles from the various government agencies, companies, and published literature from which measurement data were drawn, each measurement was assigned a job title using the Bureau of Labor Statistics' 2010 Standard Occupational Classification (SOC).[22]

The SOC structure is hierarchical and made up of major, minor, broad, and detailed groups. Figure 1 provides an example of this structure using the detailed SOC 33–9099 which corresponds to the SOC group of “Protective Service Workers, All Other” and is nested in the broad SOC 33–9090, “Miscellaneous Protective Service Workers”. The broad SOC is in turn nested in the minor SOC 33–9000, “Other Protective Service Workers,” which resides within the major SOC 33–0000, “Protective Service Occupations”.

To take advantage of the hierarchical structure of the SOC system we chose to use a parametric Bayes imputation method to impute missing values at the broad SOC level. Imputation is a widely used method for filling out missing values.[23] We performed a parametric imputation algorithm[24–26] (assuming that some broad SOC means are observed while other broad SOC means are missing at random, and that these observed broad SOC means and the broad SOC means to be estimated are all normally distributed defined by a set of parameters).[27] All models were performed in R. There were a total of 461 broad SOCs, 222 (48%) of which had missing data. Of these 222 broad SOCs four were in the major SOC group 23–0000 (Legal Occupations). Because we did not have any measurements for this occupational group we could not perform any imputation; imputation was possible for all other broad SOCs. We first created training and validation datasets to evaluate imputation accuracy by comparing observed and imputed data in the validation dataset in order to benchmark our imputation against the truth. We then used the full dataset to impute missing values for each broad SOC to be used for future research.

Model Construction and Validation

As the SOC preserves a hierarchical structure such that there is a hierarchy of nested populations, it is natural to consider using an appropriate statistical model that efficiently captures this data structure. A hierarchical model was used to estimate missing values in the dataset in our analysis.[28] The derivation of the method used is presented in Appendix 1. Let i denote the index of major SOCs and let j denote the index of broad SOCs that are nested within the major SOCs. There are two data components in this model: the observed SOCs and the missing SOCs. We assign separate indices for these two data components. For those broad SOCs that are observed, Y_{ij}^{obs} is the sample mean of the jth broad SOC in the ith major SOC. Consider a model describing our information about a hierarchical dataset $\left\{Y_1^{obs}, \ldots, Y_I^{obs}\right\}$ where $Y_i^{obs} = \left\{Y_{i1}^{obs}, \ldots, Y_{in_i}^{obs}\right\}$ consisting of all the observed data in the ith major SOC. s_{ij}^{obs} and n_{ij}^{obs} are the corresponding sample standard deviation and sample size, respectively, corresponding to the jth broad SOC nested in the ith major SOC. All that is known about this dataset are Y_{ij}^{obs}, s_{ij}^{obs} and n_{ij}^{obs} and the hierarchical structure of the dataset. θ_{ij}^{obs} is the true (unknown) mean of jth observed broad SOC in the ith major SOC and is described Equation 1 while θ_{ik}^{mis} is the true mean of kth missing broad SOC in the ith major SOC.

$$Y_{ij}^{obs} \sim N(\theta_{ij}^{obs}, \frac{(s_{ij}^{obs})^2}{n_{ij}^{obs}}) \qquad \text{Equation 1}$$

The random variables θ_{ij}^{obs} can be thought of as independent samples from the major SOC with index i, described by some fixed but unknown feature parameter θ_i and σ^2 where θ_i is the true mean of ith major SOC and σ^2 is the variation of broad SOCs within this major SOC. Similarly, the random variables θ_{ik}^{mis} can also be thought of as independent samples from the major SOC with index, i, described by θ_i and σ^2. In the normal model, we model the data as conditionally independent and identically distributed (i.i.d.) normal (θ_i, σ^2):

$$\theta_{ij}^{obs} \sim N\left(\theta_i, \sigma^2\right)$$

$$\theta_{ik}^{mis} \sim N\left(\theta_i, \sigma^2\right)$$

To represent the information about θ_i, we treat θ_i, $i = 1, \ldots, I$ as independent samples from the population mean. Assume the true population mean level is μ and the variation among all major SOCs is τ^2. Then the distribution of θ_i is:

$$\theta_i \sim N\left(\mu, \tau^2\right)$$

In sum, we have a hierarchical normal model that describes the heterogeneity of means across different broad SOCs and major SOCs. In this hierarchical model we assume that the within- and between-major SOC sampling models are both normal. We further assume that the sample mean of each broad SOC is distributed around the true mean of that broad SOC. The within-major SOC sampling variance σ^2 is assumed to be constant across major SOC groups and the between-major SOC sampling variance τ^2 is also assumed to be constant. The fixed but unknown parameters in this model are θ_{ij}^{obs}, $i = 1, \ldots, I$; $j = 1, \ldots, n_i^{obs}$, θ_{ik}^{mis}, $i = 1, \ldots, I$; $k = 1, \ldots, n_i^{mis}$, θ_i, $i = 1, \ldots, I$ and μ, τ^2, σ^2 which will be estimated. For the parameters μ, τ^2, σ^2, we need to specify prior distributions on them. We chose to use the standard conjugate normal and inverse-gamma prior distributions for these parameters as shown in equation 2.

$$\tau^2 \sim Inv - gamma\left(\frac{\eta_0}{2}, \frac{\eta_0 \tau_0^2}{2}\right); \sigma^2 \sim Inv - gamma\left(\frac{\upsilon_0}{2}, \frac{\upsilon_0 \sigma_0^2}{2}\right); \mu \sim N\left(\mu_0, \gamma_0^2\right) \qquad \text{Equation 2}$$

Implying the densities $p(\tau^2) = \frac{1}{\tau^{2(\frac{\eta_0}{2}+1)}} \exp(-\frac{\eta_0 \tau_0^2}{2\tau^2})$ and $p(\sigma^2) = \frac{1}{\sigma^{2(\frac{\upsilon_0}{2}+1)}} \exp(-\frac{\upsilon_0 \sigma_0^2}{2\sigma^2})$. Since no prior information is available, we specify non-informative priors for all these parameters. A graphical representation of the model is presented in Figure 2.

The unknown quantities include the broad SOC means θ_{ij}^{obs}, $i = 1, \ldots, I$; $j = 1, \ldots, n_i^{obs}$, θ_{ik}^{mis}, $i = 1, \ldots, I$; $k = 1, \ldots, n_i^{mis}$, the major SOC means θ_i, $i = 1, \ldots, I$, the population mean μ, the within major SOC sampling variance σ^2 and the between major SOC sampling variance τ^2. Posterior inference for these parameters can be made by constructing a Gibbs sampler, which is an iterative algorithm that construct a dependent sequence of posterior samples by sweeping through each variables to sample from its conditional distribution with the remaining variables fixed at their current values [29]. After some calculation, we find that the conditional distribution of every mean parameter, including the broad SOC means θ_{ij}^{obs}, $i = 1, \ldots, I$; $j = 1, \ldots, n_i^{obs}$, θ_{ik}^{mis}, $i = 1, \ldots, I$; $k = 1, \ldots, n_i^{mis}$, the major SOC means θ_i, $i = 1, \ldots, I$, the population mean μ, is normal. The conditional distribution of SOC sampling variance σ^2 and the conditional distribution of the between major SOC sampling variance τ^2 are both inverse gamma.

Posterior approximation proceeds by iterative sampling of each unknown quantity from its full conditional distribution. We choose the number of iterations S to be 10000 and set the starting values for each of these parameters. Given a current state of the unknowns $\left\{\theta_{11}^{obs(s)}, \ldots, \theta_{In_I}^{obs(s)}, \theta_{11}^{mis(s)}, \ldots, \theta_{In_I}^{mis(s)}, \theta_i^{(s)}, \mu^{(s)}, \tau^{2(s)}, \sigma^{2(s)}\right\}$, a new state is generated as follows:

1. Posterior step: sample $\theta_i^{(s+1)}$, $i = 1, \dots, I$ from $\theta_i \mid \mu^{(s)}, \theta_{i1}^{obs(s)}, \dots, \theta_{in_i}^{obs(s)}, \theta_{i1}^{mis(s)}, \dots, \theta_{in_i}^{mis(s)}, \tau^{2(s)}, \sigma^{2(s)}$ based on its full conditional distribution

2. Posterior step: sample $\mu^{(s+1)}$ from $\mu \mid \theta_1^{(s+1)}, \dots, \theta_I^{(s+1)}, \tau^{2(s)}$

3. Posterior step: sample $\tau^{2(s+1)}$ from $\tau^2 \mid \theta_1^{(s+1)}, \dots, \theta_I^{(s+1)}, \mu^{(s+1)}$

4. Posterior step: sample $\sigma^{2(s+1)}$ from $\sigma^2 \mid \theta_{11}^{obs(s)}, \dots, \theta_{In_I}^{obs(s)}, \theta_{11}^{mis(s)}, \dots, \theta_{In_I}^{mis(s)}, \theta_1^{(s+1)}, \dots, \theta_I^{(s+1)}$

5. Posterior step: sample $\theta_{ij}^{obs(s+1)}$, $i = 1, \dots, I$, $j = 1, \dots, n_i^{obs}$ from $\theta_{ij}^{obs} \mid \theta_i^{(s+1)}, \sigma^{2(s+1)}$

6. Imputation step: sample $\theta_{ij}^{mis(s+1)}$, $i = 1, \dots, I$, $j = 1, \dots, n_i^{mis}$ from $\theta_{ij}^{mis} \mid \theta_i^{(s+1)}, \sigma^{2(s+1)}$

The procedures were repeated S times until convergence has reached. After a thinning procedure and a burn-in period, the draws were used for the posterior inference. A detail description of this Bayesian parametric imputation procedure is presented in Appendix 1.

Prior to imputation of the full JEM, the imputation model was evaluated by dividing the available data in to a training and validation set. The training dataset consisted of 189 broad SOCs that were randomly chosen from the available dataset of 239 broad SOCs provided the broad SOC contained more than one measurement, as imputation cannot be conducted with one measurement. The remaining 50 broad SOCs, including those with a single measurement, were assigned to the validation dataset. The posterior distribution of the mean and variances was calculated at the broad and major SOC level in the training dataset and compared to the observed data in the validation dataset. After the model evaluation, the training and validation datasets were combined, and all data were used for imputation of the final JEM. A level of confidence was assigned for each estimate based on the width of that estimate's 95% creditable interval. Estimates with a 95% creditable interval with a width <3 dB were considered high confidence, ≤3 dB but ≤12 dB moderate confidence, and >12 dB low confidence. These values were chosen because an increase of 3 dB roughly equivalent to doubling sound power and is also the doubling rate used by the US National Institute for Occupational Safety and Health (NIOSH), European Union, and International Organization for Standardization. [30–32]

Temporal changes in exposure patterns have been shown to be important for multiple different agents. [14,20,33] However, the scarcity of data in certain broad level SOCs made it impractical to include a factor for the effect of time in the imputation model, in our study. Considering the possibility of temporal trend, a sub-analysis was further conducted to determine the effect of time on noise exposure levels across all the major SOCs. We chose five different year bins (before 1984, 1984–1992, 1993–2000, 2001–2009, and after 2009) which are approximately equal in length and also reflect regulatory changes promulgated by OSHA and MSHA. This analysis cannot be used to adjust the estimates from the main analysis but provides additional insight in to a possible source of error in in the exposure estimates.

Code Availability

Both the STATA and the R code used for this analysis is available upon request of the corresponding authors.

Results

A summary of the estimates from the model validation is presented in Table 2, where the population mean (μ), is estimated to be 82.4 dBA, the within-major SOC variance (σ^2) is 20.0 and the between-major SOC variance (τ^2) is 13.3. The estimated mean noise exposure for each major SOC ranged from 78.4 (43–0000, "Office and Administrative Support Occupations") to 85.5 dBA (45–0000. "Farming, Fishing, and Forestry Occupations"). The 95% credible interval varied depending on the number of broad SOCs present within each major SOC (Table 3). Figure 3. displays a fairly strong agreement between the 189 estimated and observed broad SOC means in the training dataset. However, Figure 3b illustrates that the agreement between the observed and predicted SOC means in the validation dataset was not as strong as the training dataset as expected. Of the 50 broad SOCs in the validation dataset 11 observed sample means were outside the 95% credible interval and 39 fell inside the credible interval, however, 7 of those broad SOCs that fell outside contained only one measurement (Figure 4).

Table 4 summarizes the population mean, and the within- and between-major SOC variance for the entire dataset (i.e. the combined validation and training datasets). The population mean was estimated to be 82.1 dBA and the within- and between-major SOC variance was estimated to be 22.1 and 13.8, respectively. The estimated mean noise exposure for each major SOC ranged from 78.6 (25–0000, "Education, Training, and Library Occupations") to 86.4 dBA (45–0000, "Farming, Fishing, and Forestry Occupations"). Similar to what we observed in the model validation results (Table 3), major SOCs that consisted of a larger number of broad SOCs had smaller 95% credible intervals.

The model predictions at the broad SOC level can be found in Appendix 2 or online at (http://noisejem.sph.umich.edu/full_results.pdf). The estimated population mean was 82.1 dBA while the estimated population standard deviation was 3.1 dBA. Of the 443 broad SOCs, 338 (76.3%) were found to have an estimated mean exposure >80 dBA, while 85 (19.2%) were found to have an estimated mean exposure greater than the current OSHA AL of 85 dBA. Additionally, 10 broad SOCs were found to have an estimated mean exposure greater that the OSHA PEL of 90 dBA. The distribution of estimated broad SOC means can be found in Figure 5, which indicates that the majority of broad SOCs have estimated mean noise exposure levels between 80 and 85 dBA. A total of 99 (22.3%) and 108 (24.3%) of the broad SOCs were found to have a high and moderate level of confidence in the estimate respectively.

An additional sub-analysis was conducted, attempting to determine the effect of time on exposure estimates. The results of this additional analysis found that nine (40.9%) of the major SOCs (11–0000, 25–0000, 37–0000, 41–0000, 43–0000, 47–0000, 49–0000, 51–0000, and 53–0000) had decreasing exposures over time. This suggests that for some broad and major SOCs temporal trends may impact exposures estimates. While these results provide additional insight in regard to the impact of time on the original exposure estimate, these new results have less practical use because the major SOCs do not provide sufficient job title specificity to accurately assign exposure estimates. The full results of this additional analysis can be found in Appendix 3.

Discussion

In this study we used principled validation strategy to evaluate the performance of an imputation strategy to estimate noise exposures in a large JEM. The imputation strategy borrows information across broad SOCs by assuming a common hierarchical distribution with parameters that are shared. The imputed SOC means were assessed for imputation accuracy in a validation dataset consisting of randomly chosen subset of SOCs. The strong agreement between the 189 estimated and observed broad SOC means in the training dataset occurred because these observed broad SOCs were used to build the hierarchical model and thus their data were “known” to the model, which yielded statistically overly optimistic estimates. The broad SOCs in the validation dataset were not used in building the hierarchical model and were thus “unknown”. The estimated SOC mean of a broad SOC in the training set was a weighted average of the observed SOC mean Y_{ij}^{obs} and the estimate of major SOC mean θ_i that it was nested in, and the weights were proportional to the estimated σ^2 (variation within major SOC) and $\frac{\left(s_{ij}^{obs}\right)^2}{n_{ij}^{obs}}$ (variation in the observed SOC mean). As the variation within major SOCs was high and the variation in the observed broad SOC means were small for most broad SOCs, the estimated broad SOC mean would be more similar to the observed broad SOC mean than the major SOC mean, if that broad SOC mean had been observed. However, the estimated mean of a broad SOC in the validation set was entirely based on the estimated mean of the major SOC that it was nested in; no additional information was available that could be used for this purpose. As a result, the agreement between the observed and predicted SOC means in the validation dataset were not as strongly associated as the training dataset.

Our estimates were developed from large datasets of measurements provided by the government, private industry, and the published literature. By taking advantage of the hierarchical structure of the SOC system we were able to use imputation to iteratively impute the missing values of the mean of the broad SOCs and to draw updated samples of the parameters based on both the means of the observed broad SOCs and the means of the missing broad SOCs. Due to the limited sample size within each minor SOC, we chose to ignore the minor SOC level in this hierarchical model. Instead we assumed that the broad SOCs within the same major SOC are more alike those broad SOCs in other major SOCs. However, if more data are available in the future and there are at least moderate numbers of broad SOC with observed measurements for most minor SOCs, it is possible to construct a hierarchical model with major SOC level, minor SOC level and broad SOC level. Such a hierarchical model incorporating the minor SOC level may be able to provide more accurate estimates of the broad SOC means. The validation analysis on the 50 randomly chosen SOCs provide a realistic sense of accuracy when a new missing exposure is predicted for an SOC. The level of confidence assigned to each estimate indicated that 236 (53.3%) of the broad SOCs had a 95% creditable interval wider than 12 dBA which suggests that caution should be exercised when using these exposure estimates until additional data can be collected, or the current estimates can be validated.

In the parametric Bayes imputation method that we used, we plugged in the posterior mean estimates of the unknown quantities as our single imputation results. However instead we could possibly create random draws from the posterior distributions of these quantities and then create multiple imputed datasets. The advantage of multiple imputation over the single imputation is that it takes into account the uncertainty in the imputation procedure.

Another potential source of error in our exposure estimates occurs because these data represents occupational noise exposures from 1970–2014. As reported by Middendorf in 2004 and Roberts et al. in 2016 occupational noise exposures have been decreasing overall in the general industry and mining sectors.[14,20] However, the results of the few other longitudinal analyses of occupational noise exposures suggest that workers in the construction and manufacturing industries may not have experienced significant reductions over time. [34,33] If a majority of measurements for a particular occupation were clustered in a short time span then it is possible that the measurements used by the model to develop exposure estimates may be biased.

The largest potential source of error in our estimates is likely the variability of exposure within each broad SOC. This is a common issue for any JEM that attempts to quantify exposures across several different industries. As identified by Rappaport et al. there is considerable variation in personal exposure for workers with similar job titles within the same workplace.[15] Grouping workers by job title is common practice in industrial hygiene because it is easy and straightforward to assign workers to an occupational group. However, as Anderson et al. have demonstrated, the standard occupational coding systems used in Canada were inadequate to accurately group workers in the pulp and paper industry.[35]

We recognize that these shortcomings of the SOC system may result in misclassification of exposure. This misclassification can be exasperated by the model when limited data is available for a broad SOC within a major SOC where other broad SOCs with dissimilar exposures influence the major SOC mean. For example, the model estimated that the broad SOC 11–1010 (Chief Executives) had a mean exposure of 84.8 dBA, which runs counter to most professional intuitions. However, this high exposure value is due in part to the fact that the major SOC 11–0000 (Management Occupations) contains broad SOCs for jobs such as “Industrial Production Managers” and “Farmers, Ranchers, and other Agricultural Managers” who would be expected to have higher exposures and thus influence the exposure estimate for the “Chief Executive” broad SOC. This is due in part, to the fact that the SOC system was designed to track economic indicators and was not intended as a classification scheme for forming similar exposure groups. However, it is still advantages to use this system, as there are numerous crosswalks available to convert SOC codes to other occupational classifications systems so that the exposure estimates can be more easily used in epidemiological studies.

However, the variability of broad SOC mean would be expected to decrease as the number of measurements increase because it would be expected that as more measurements are added to a broad SOC that the estimated mean would become closer to the true mean of the broad SOC. Exposure estimates could be further enhanced by using more informative priors based on expert knowledge and information from the published literature. However, we chose to make the imputation process more robust and less sensitive to subjective choices at the cost of making the process less efficient. Future efforts will be focused on incorporating expert judgment to enhance the accuracy of the JEM's estimates, particularly for broad SOCs that we had low levels of confidence in the estimates.

The results of our analysis indicated that the majority of broad SOCs were estimated to be exposed to noise ≥ 80.0 and <85.0 dBA. While these broad SOCs are not estimated to exceed the OSHA action level, it is worth noting that the average estimated exposure and standard deviation for broad SOCs in this group were 82.3 and 3.6 dBA, respectively, with a 95% confidence interval between 72.3 and 89.4 dBA. This suggests that while the estimated mean exposure for these groups was below the action level, there is considerable variability in these exposures that must be considered when using these estimates to identify occupations that should be enrolled in hearing conservation programs (HCPs). In other words, individual exposures or minor SOCs within the broad SOC groups in the ≥ 80.0 and <85.0 dBA bin may still exceed the action level. This is in contrast to broad SOCs that are in the >=85.0, <90.0 dBA and > 90.0 dBA groups, which have an average estimated exposure of 87.1, 91.6 dBA and standard deviations of 1.2 and 0.8 dBA, respectively. For these two groups, there is far greater confidence that noise exposures exceed the action level or PEL and that location-specific measurements should be taken to determine if controls should be implemented to protect workers from excessive exposure.

Exposure estimates for individual broad SOCs can be found in Appendix 2. While these estimates cannot replace personal measurement data, they do provide a starting point for occupational health professionals to identify workers who may be overexposed to noise. Additionally, the provided measure of variability will help inform and guide the decisions of occupational health professionals regarding workers in job groups whose exposure may vary from day to day depending on the specific work tasks being conducted. Note that these exposure estimates are calculated based on the currently available data in the JEM. If new measurements are added to the JEM in the future, these exposure estimates can be refined and updated.

To our knowledge the exposure estimates from our model are based on the most comprehensive dataset of occupational noise exposure ever collected. The only other instance of a comprehensive JEM developed for occupational noise was reported by Sjöström et al. in 2013. The authors of that paper used a mixture of 569 quantitative noise measurements and qualitative measurements made by expert judgment to assign exposure groupings for 129 unique job families.[13] In contrast to what has been seen in the US, occupational noise exposures in Sweden saw only a slight decrease from 1970 to 2004 which, likely reflects the difference in the dates of promulgation and enforcement of occupational health laws in the US compared to Sweden.[13,14] It is not straightforward to directly compare the results from our JEM to the JEM constructed by Sjöström et al. because we only used quantitative measurements in our JEM. In addition, Sweden uses a more protective noise exposure standard than OSHA (85 dBA criterion level and 3 dB time-intensity exchange rate) while OSHA uses the less protective 90 dBA criterion level and 5 dB time-intensity exchange rate, making it impossible to directly compare the measurements.[36]

Despite the limitations associated with this JEM we believe it represents a useful tool for occupational health professionals and researchers. Our future plans include combining the exposure estimates from this model with information on the frequency of noise exposure from Department of Labor’s Occupational Information Network (O*NET) system by using responses from survey question 4.C.2.b.1.a, which asks respondents to provide a response from 0–100% “How often does this job require working exposed to sounds and noise levels that are distracting or uncomfortable?”.[37] This will build on previous work by Choi et al. that used the responses from O*NET’s databases to create statistical models to predict NIHL.[38] Our exposure estimates can also be used with noise-induced hearing loss models published by the International Organization for Standards (ISO) to predict hearing threshold levels of participants in the National Health and Nutrition Examination Survey (NHANES) which contains both audiometric and employment history data.[39,40] Finally, the estimates in our JEM may be used to drive additional targeted surveillance and assessment efforts in specific occupations; these efforts could leverage smart device-based measurement technologies, which under certain circumstances can yield low-cost, reasonably accurate noise exposure measurements.[41,42] Each of these steps will yield better noise exposures estimates that can, in turn, be used to guide efforts to control noise exposures and reduce occupational NIHL.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by the National Institute of Occupational Safety and Health Grant # R21OH0 10482: Development of a US/Canadian Job Exposure Matrix (JEM) for Noise

References

1. Tak S, Calvert GM. Hearing difficulty attributable to employment by industry and occupation: an analysis of the National Health Interview Survey--United States, 1997 to 2003. J Occup Environ Med 2008; 50: 46–56.
2. Saunders G, Griest S. Hearing loss in veterans and the need for hearing loss prevention programs. Noise Heal 2009; 11: 14–21.
3. Themann C, Suter AH, Stephenson MR. National research agenda for the prevention of occupational hearing loss-part 1. Semin Hear 2013; 34: 145–207.

4. Neitzel RL, Swinburn TK, Hammer MS, Eisenberg D. Economic Impact of Hearing Loss and Reduction of Noise-Induced Hearing Loss in the United States. J speech, Lang Hear Res 2017; 25: 1–8.

5. Tak S, Davis RR, Calvert GM. Exposure to hazardous workplace noise and use of hearing protection devices among US workers--NHANES, 1999–2004. Am J Ind Med 2009; 52: 358–71. [PubMed: 19267354]

6. NIOSH. Criteria for a Recommended Standard Occupational Noise Exposure Revised Criteria 1998. National Institutes of Occupational Safety and Health, 1998 doi: 98–126.

7. Seixas NS, Checkoway H. Exposure assessment in industry specific retrospective occupational epidemiology studies. Occup. Environ. Med. 1995; 52: 625–33. [PubMed: 7489051]

8. Dewar R, Siemiatycki J, Gerin M. Loss of Statistical Power Associated with the Use of a Job-Exposure Matrix in Occupational Case-Control Studies. Appl Occup Environ Hyg 1991; 6: 508–515.

9. Astrakianakis G, Anderson JTL, Anderson JTL, Keefe AR, Bert JL, Le N et al. Job—exposure matrices and retrospective exposure assessment in the pulp and paper industry. Appl Occup Environ Hyg 1998; 13: 663–670.

10. Friesen MC, Demers PA, Spinelli JJ, Le ND. Validation of a semi-quantitative job exposure matric at an aluminum Smelter. Ann Occup Hyg 2003; 47: 477–484. [PubMed: 12890656]

11. Guéguen a, Goldberg M, Bonenfant S, Martin JC. Using a representative sample of workers for constructing the SUMEX French general population based job-exposure matrix. Occup. Environ. Med. 2004; 61: 586–93. [PubMed: 15208374]

12. Semple SE, Dick F, Cherrie JW. Exposure assessment for a population-based case-control study combining a job-exposure matrix with interview data. Scand J Work Environ Health 2004; 30: 241–248. [PubMed: 15250653]

13. Sjöström M, Lewné M, Alderling M, Willix P, Berg P, Gustavsson P et al. A job-exposure matrix for occupational noise: development and validation. Ann Occup Hyg 2013; 57: 774–83. [PubMed: 23380283]

14. Middendorf PJ. Surveillance of occupational noise exposures using OSHA’s Integrated Management Information System. Am J Ind Med 2004; 46: 492–504. [PubMed: 15490475]

15. Rappaport SM, Kromhout H, Symanski E. Variation of exposure between workers in homogeneous exposure groups. Am Ind Hyg Assoc J 1993; 54: 654–662. [PubMed: 8256689]

16. Stewart PA, Herrick RF. Issues in Performing Retrospective Exposure Assessment. Appl Occup Environ Hyg 1991; 6: 421–427.

17. OSHA. OSHA Technical Manual Noise. 2013https://www.osha.gov/dts/osta/otm/new_noise/index.html.

18. Cheng W, Roberts B, Mukherjee B, Neitzel RL. Meta-Analysis of Job Exposure Matrix Data from Multiple Sources. J Expo Sci Environ Epidemiol 2017; In Press. doi:10.1038/jes.2017.19.

19. MSHA. Subchapter M- Uniform Mine Health Regulations Part 62 - Occupational Noise. United States, 2014.

20. Roberts B, Sun K, Neitzel RL. What can 35 years and over 700,000 measurements tell us about noise exposure in the mining industry? Int J Audiol 2016; 2027. doi:10.1080/14992027.2016.1255358.

21. Office of Management Budget. North American Industry Classification System; Revision for 2012; Notice. US Congress: Washington, D.C., 2011.

22. Bureau of Labor Statistics. 2010 SOC User Guide. Alexandria, VA, 2010http://www.bls.gov/soc/soc_2010_user_guide.pdf.

23. Chen D-G, Chen J, Lu X, Yi G, Yu H (eds.). Advanced Statistical Methods in Data Science. Wiley: New York, NY, 2016.

24. Little R, Rubin D. Statistical Analysis with Missing Data. 2nd ed. Wiley: New York, NY, 2002.

25. Rubin D Multiple Imputation in Sample Surveys- A Phenomenological Bayesian Approach to Nonresponse. In: Proceedings of the survey reseach methods section. American Statistical Association: Washington DC, 1978, pp 20–34.

26. Rubin D Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons: New York, NY, 1987.
27. Schenker N, Taylor JMG. Partially parametric techniques for multiple imputation. Comput Stat Data Anal 1996; 22: 425–446.
28. Hoff PD. A First Course in Bayesian Statistical Methods. Springer-Verlag: New York, NY, 2009.
29. Gilks W, Richardsons S, Spiegelhalter D. Markov Chain Monte Carlo in Practice. Chapman and Hall: London, United Kingdom, 1996.
30. NIOSH. Occupational Noise Exposure Criteria for a Recommended Standard. Cincinnati, Ohio, 1998 doi:DHHS (NIOSH) Publication No. 98–126.
31. ISO. Acoustics - Determination of occupational noise exposure - Engineering method. Geneva, Switzerland, 2008.
32. Lambert J, Lelong J, Phillips-Bertin C. Final Report ENNAH – European Network on Noise and Health. ENNAH -European Netw Noise Heal 2013; : 178.
33. Neitzel R, Galusha D, Dixon-Ernsts C, Rabinowitz P. Methods for evaluating temporal trends in noise exposure. Int J Audiol 2014; 53: S76–S83.
34. Neitzel RL, Stover B, Seixas NS. Longitudinal assessment of noise exposure in a cohort of construction workers. Ann Occup Hyg 2011; 55: 906–16. [PubMed: 21825303]
35. Anderson JTL, Astrakianakis G, Band PR. Standardizing job titles for exposure assessment in the pulp and paper industry. Appl Occup Environ Hyg 1997; 12: 611–614.
36. Berger EH, Royster LH, Royster JD, Driscoll DP, Layne M (eds.). The Noise Manual. 5th ed. American Industrial Hygiene Association: Fairfax, VA, 2003.
37. O*NET OnLine. National Center for O*NET Development. Work Context: Sounds, Noise Levels Are Distracting or Uncomfortable. https://www.onetonline.org/find/descriptor/result/4.C.2.b.1.a (accessed 9 Jan2017).
38. Choi Y-H, Hu H, Tak S, Mukherjee B, Park SK. Occupational noise exposure assessment using O*NET and its application to a study of hearing loss in the US general population. Occup Environ Med 2012; 69: 176–183. [PubMed: 21725070]
39. National Center for Health Statistics. National Health and Nutrition Examination Survey 1999 – 2014 Survey Content Brochure. 2014.
40. ISO. Acoustics - Estimation of noise-induced hearing loss. Geneva, Switzerland, 2013.
41. Roberts B, Kardous C, Neitzel RL. Improving the Accuracy of Smart Devices to Measure Noise Exposure. J Occup Environ Hyg 2016; 0: 0.
42. Kardous C, Shaw PB. Evaluation of smartphone sound measurement applications. J Acoust Soc Am 2014; 135: EL186–EL192. [PubMed: 25236152]

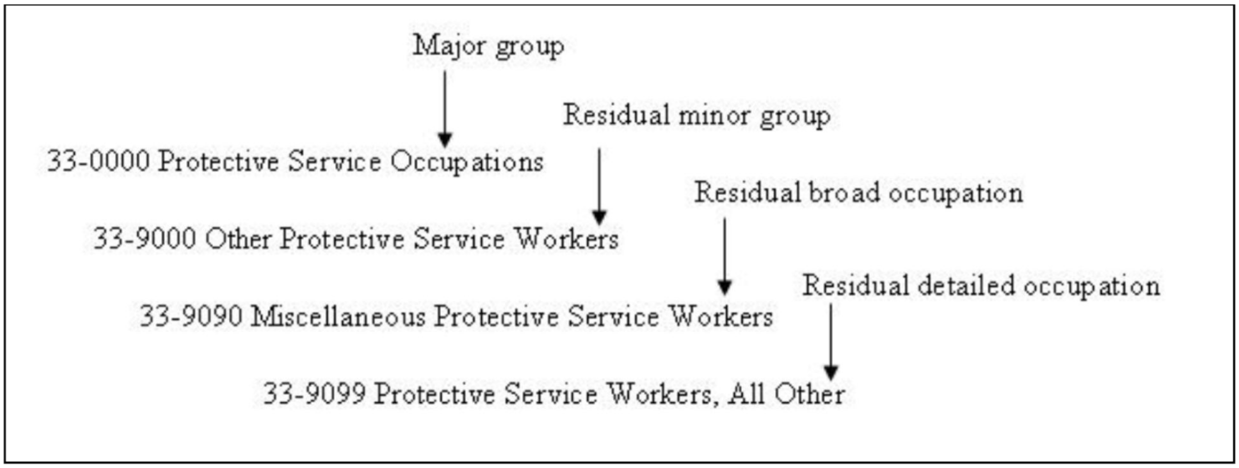

Major group
33-0000 Protective Service Occupations
Residual minor group
33-9000 Other Protective Service Workers
Residual broad occupation
33-9090 Miscellaneous Protective Service Workers
Residual detailed occupation
33-9099 Protective Service Workers, All Other

Figure 1.

Example of the hierarchical structure in the SOC system reprinted from the 2010 SOC User Guide (22).

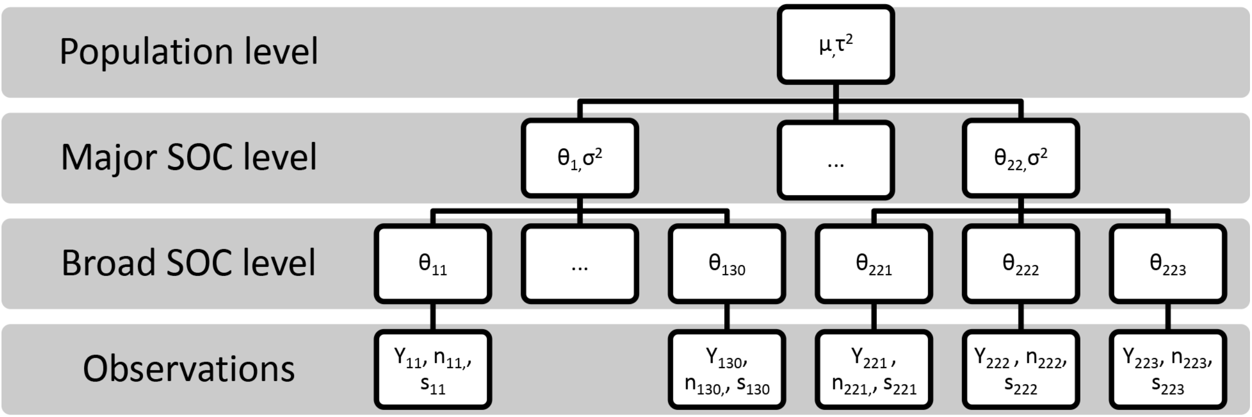
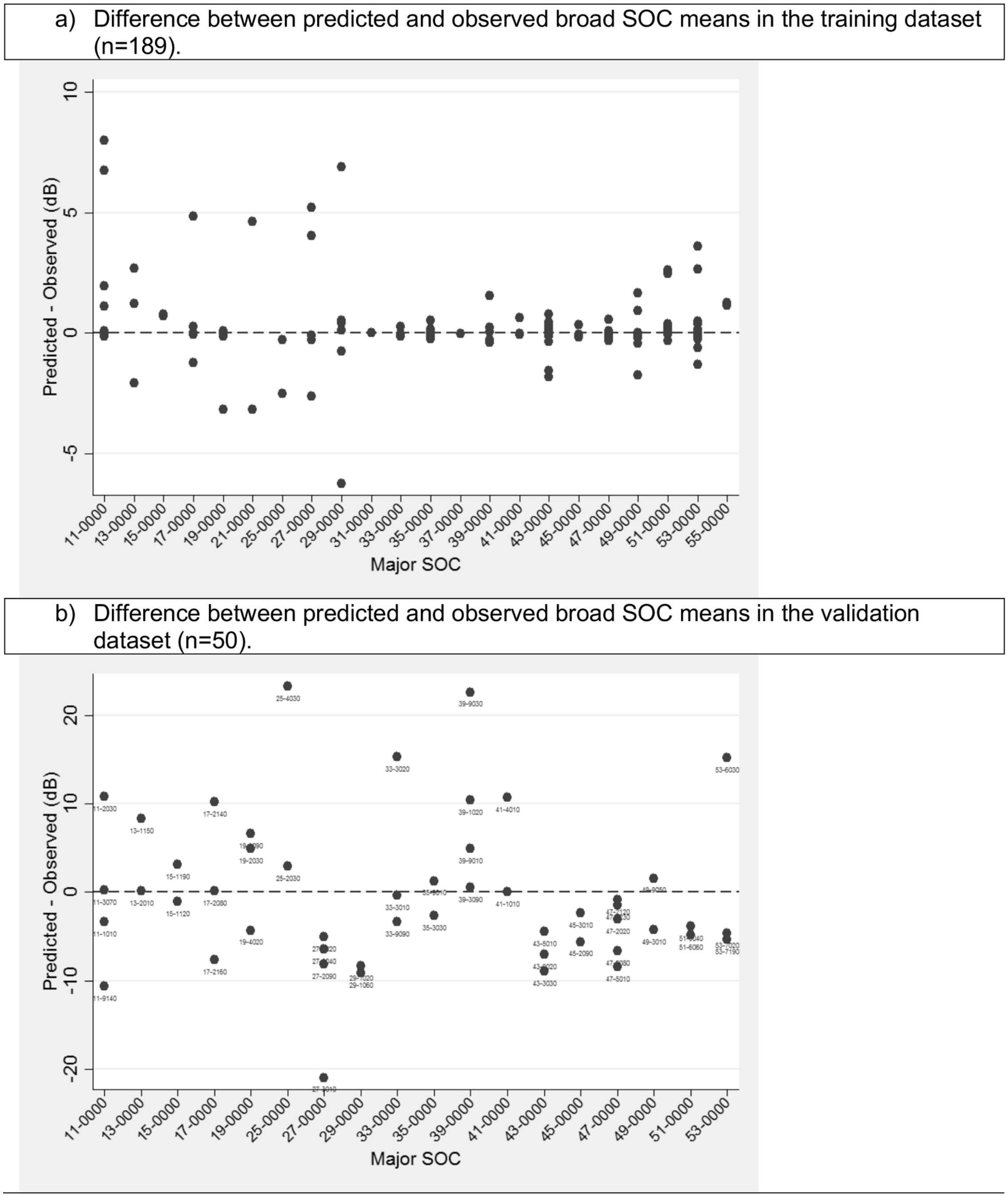


Figure 2.

An illustration of the hierarchical structure used in this analysis. There are 22 major SOCs and various number of broad SOCs within each major SOC. For example, the first major SOC has 22 broad SOCs and the 22nd major SOC has 3 broad SOCs.



a) Difference between predicted and observed broad SOC means in the training dataset (n=189).

b) Difference between predicted and observed broad SOC means in the validation dataset (n=50).

Figure 3.

Comparison of predicted and observed broad SOC means for the training (a) and validation (b) dataset set.

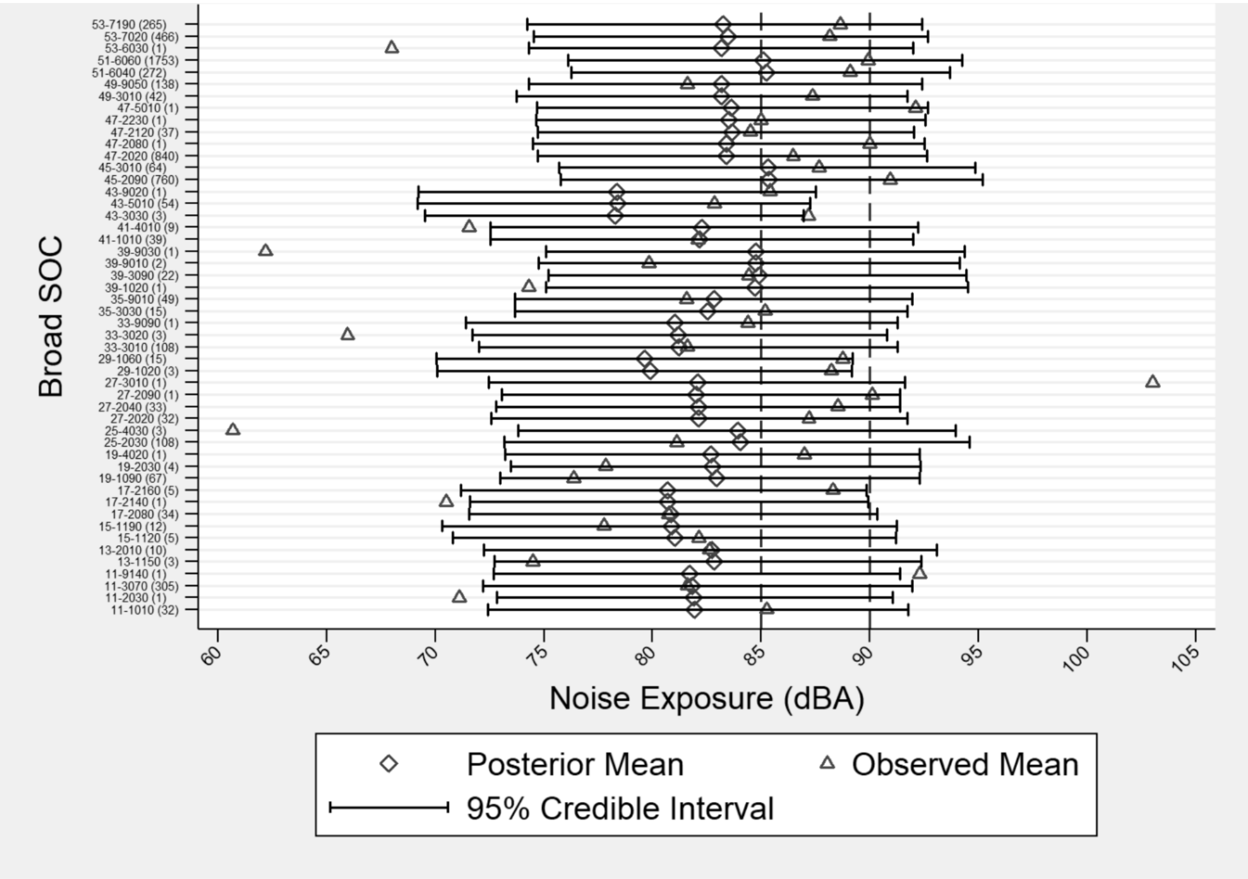


Figure 4.
Posterior and observed broad SOCs means for the validation dataset (n=50). The sample size for the observed mean is shown in parentheses.

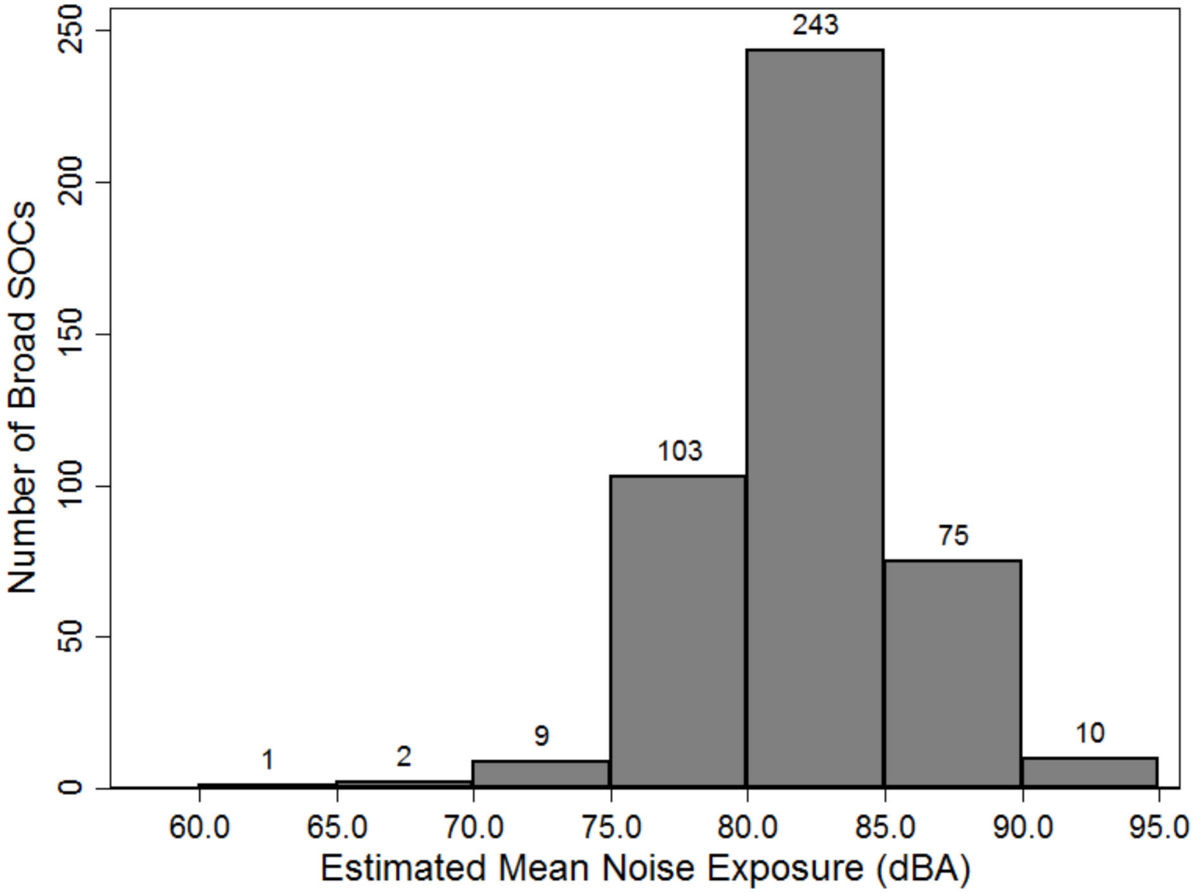


Figure 5.

The distribution of estimated mean noise exposures (dBA) at the broad SOC level. The numbers above the bars indicate the number of SOCs with estimated mean exposures that lie within that bar.

Table 1.

Source of data used in the analysis.

Total	Before 1984	1984–1992	1993–2000	2001–2009	After 2009	Total
Total						
	109 123	203 071	157 471	198 987	85 050	753 702
Government						
MSHA	90 305	187 886	142 899	151 078	70 909	643 077
OSHA	55 318	159 701	132 302	135 753	66 065	549 139
	34 987	28 185	10 597	15 325	4 844	93 938
Private Industry						
Agriculture Forestry Fishing and Hunting	18 256	15 067	13 566	46 107	11 922	104 918
Mining Quarrying and Oil and Gas Extraction	638	44	56	15	51	804
Utilities	1 388	6 085	5 695	7 514	1 830	22 512
Construction	13	8	111	56	458	646
Manufacturing	827	186	719	1 113	161	3 006
Wholesale Trade	14 995	8 279	6 272	36 875	8 986	75 407
Retail Trade	16	17	4	194	129	360
Transportation and Warehousing	30	12	87	13	25	167
Information	38	152	56	51	96	393
Finance and Insurance	26	11	1	1	0	39
Real Estate and Rental and Leasing	0	0	0	0	0	0
Professional Scientific and Technical Services	0	0	0	0	1	1
Management of Companies and Enterprises	9	0	2	11	1	23
Administrative and Support and Waste Management and Remediation Services	0	0	0	0	0	0
Educational Services	19	14	11	2	8	54
Health Care and Social Assistance	42	17	39	124	125	347
Arts Entertainment and Recreation	26	14	44	18	8	110
Accommodation and Food Services	0	0	11	11	9	31
Other Services (except Public Administration)	6	28	34	46	2	116
Public Administration	67	69	194	56	14	400
	116	131	230	7	18	502
Published Literature						
	562	118	1 006	1 802	2 219	5 707

Table 2.

Summary of posterior distribution of parameters from the model validation.

Parameter	Posterior mean	Posterior standard deviation	95% Credible interval
μ	82.3	0.9	80.6–84.2
σ^2	20.0	2.5	15.7–25.9
σ	4.4	0.3	3.9–5.1
τ^2	13.3	5.3	6.2–26.5
τ	3.5	0.7	2.5–5.2

Table 3.

Posterior distribution of major SOC means from the model validation.

Major SOC	Major SOC Title	Posterior mean	Posterior standard deviation	95% credible interval	Number of broad SOCs[1]	Total Number of Measurements
11–0000	Management Occupations	81.8	1.8	78.4–85.3	7	277
13–0000	Business and Financial Operations Occupations	82.7	2.4	78.2–87.6	3	39
15–0000	Computer and Mathematical Occupations	80.9	2.7	75.4–86.1	2	25
17–0000	Architecture and Engineering Occupations	80.7	1.6	77.6–84.0	7	1 446
19–0000	Life, Physical, and Social Science Occupations	82.8	2.0	78.9–86.8	4	183
21–0000	Community and Social Service Occupations	80.7	2.8	74.7–86.0	2	7
25–0000	Education, Training, and Library Occupations	84.0	2.9	78.5–89.6	2	33
27–0000	Arts, Design, Entertainment, Sports, and Media Occupations	82.1	2.00	78.2–86.1	5	77
29–0000	Healthcare Practitioners and Technical Occupations	79.9	1.8	76.2–83.3	6	89
31–0000	Healthcare Support Occupations	82.3	2.9	76.6–87.9	1	15
33–0000	Protective Service Occupations	81.2	1.8	77.6–84.7	5	106
35–0000	Food Preparation and Serving Related Occupations	82.7	1.7	79.7–85.9	8	107
37–0000	Building and Grounds Cleaning and Maintenance	85.0	2.5	80.2–89.8	2	353
39–0000	Personal Care and Service Occupations	84.8	1.9	80.9–88.6	5	47
41–0000	Sales and Related Occupations	82.3	2.1	78.2–86.6	3	191
43–0000	Office and Administrative Support Occupations	78.4	1.2	76.2–80.6	16	433
45–0000	Farming, Fishing, and Forestry Occupations	85.5	2.0	81.7–89.5	4	305
47–0000	Construction and	83.5	0.9	81.8–85.1	27	93 531
49–0000	Extraction Occupations Installation, Maintenance, and Repair Occupations	83.3	1.2	80.9–85.5	14	8 923
51–0000	Production Occupations	85.2	0.7	83.9–86.6	43	26 989
53–0000	Transportation and Material Moving Occupations	83.3	0.9	81.5–85.2	21	16 456
55–0000	Military Specific Occupations	78.9	2.8	73.2–83.9	2	12

[1] Number of broad SOCs in the training dataset

Table 4.

Summary of posterior distribution of parameters from the model imputation.

Parameter	Posterior mean	Posterior standard deviation	95% Credible interval
μ	82.1	0.9	80.3–83.9
σ^2	22.1	2.5	17.7–27.5
σ	4.7	0.3	4.2–5.3
τ^2	13.8	5.1	6.6–26.6
τ	3.7	0.7	2.6–5.2

Table 5.

Posterior distribution of major SOC means from the model imputation.

Major SOC	Major SOC Title	Posterior mean	Posterior standard deviation	95% credible interval	Number of broad SOCs[1]	Total Number of Measurements
11–0000	Management Occupations	82.0	1.6	78.6–85.1	9	1 380
13–0000	Business and Financial Operations Occupations	81.4	2.0	77.3–85.1	5	39
15–0000	Computer and Mathematical Occupations	80.4	2.3	75.9–84.8	4	25
17–0000	Architecture and Engineering Occupations	81.3	1.5	78.3–84.4	9	7 176
19–0000	Life, Physical, and Social Science Occupations	81.4	1.9	77.7–85.1	6	776
21–0000	Community and Social Service Occupations	80.6	3.0	74.7–86.3	2	7
25–0000	Education, Training, and Library Occupations	78.6	2.2	74.1–82.9	4	139
27–0000	Arts, Design, Entertainment, Sports, and Media Occupations	83.5	1.9	79.9–87.1	7	264
29–0000	Healthcare Practitioners and Technical Occupations	81.5	1.7	78.2–84.9	8	220
31–0000	Healthcare Support Occupations	82.1	2.9	76.4–87.9	1	67
33–0000	Protective Service Occupations	79.7	1.6	76.5–82.9	7	480
35–0000	Food Preparation and Serving Related Occupations	82.8	1.42	79.9–85.6	10	319
37–0000	Building and Grounds Cleaning and Maintenance	84.6	2.6	79.7–89.8	2	1 675
39–0000	Personal Care and Service Occupations	84.6	1.8	81.0–88.2	7	178
41–0000	Sales and Related Occupations	81.1	1.9	77.4–84.8	5	935
43–0000	Office and Administrative Support Occupations	78.8	1.1	76.6–80.9	18	2 038
45–0000	Farming, Fishing, and Forestry Occupations	86.4	1.8	83.0–89.8	6	1 384
47–0000	Construction and Extraction Occupations	83.6	0.9	81.9–85.3	29	46 9231
49–0000	Installation, Maintenance, and Repair Occupations	83.3	1.2	81.2–85.7	16	44 769
51–0000	Production Occupations	85.4	0.7	84.0–86.8	45	135 533
53–0000	Transportation and Material Moving Occupations	83.7	1.0	81.8–85.7	23	81 951
55–0000	Military Specific Occupations	78.8	2.8	73.1–84.1	2	12

[1] Total number of broad SOCs in the training and validation datasets