*Annals of Work Exposures and Health,* 2018, Vol. 62, No. 3, 269–280
doi: 10.1093/annweh/wxy003
Advance Access publication 3 February 2018
Original Article



Original Article

# Characterization of the Selective Recording of Workplace Exposure Measurements into OSHA's IMIS Databank

**Philippe Sarazin[1,2]*, Igor Burstyn[3], Laurel Kincl[4], Melissa C. Friesen[5] and Jérôme Lavoué[2,6]**

[1]Chemical and Biological Hazards Prevention, Institut de recherche Robert-Sauvé en santé et en sécurité du travail, Montréal, Québec, Canada; [2]Department of Occupational and Environmental Health, Université de Montréal, Montréal, Québec, Canada; [3]Environmental and Occupational Health, Drexel University, Philadelphia, Pennsylvania, United States; [4]College of Public Health and Human Sciences, Oregon State University, Corvallis, Oregon, United States; [5]Division of Cancer Epidemiology & Genetics, Occupational and Environmental Epidemiology, National Cancer Institute, Rockville, Maryland, United States; [6]University of Montreal Hospital Research Centre, Montréal, Québec, Canada

*Author to whom correspondence should be addressed. Tel.: +(514) 288–1551 (ext. 402); fax: +1-514-288-7446;
e-mail: philippe.sarazin@irsst.qc.ca

Submitted 20 March 2017; revised 8 November 2017; editorial decision 12 December, 2017; revised version accepted 9 January 2018.

## Abstract

**Objectives:** The Integrated Management Information System (IMIS) is the largest multi-industry source of exposure results available in North America. In 2010, the Occupational Safety and Health Administration (OSHA) released the Chemical Exposure Health Data (CEHD) that contains analytical results of samples collected by OSHA inspectors. However, the two databanks only partially overlap, raising suspicion of bias in IMIS data. We investigated the factors associated with selective recording of CEHD results into the IMIS databank.

**Methods:** This analysis was based on personal exposure measurements of 24 agents from 1984 to 2009. The association between nine variables (level of exposure coded as detected versus non-detected (ND), whether a sampling result was part of a panel of chemicals, duration of sampling, issuance of a citation, presence of other detected levels during the same inspection, year, OSHA region, amount of penalty, and establishment size) and a CEHD sampling result being reported in IMIS was analyzed using modified Poisson regression.

**Results:** A total of 461 900 CEHD sampling results were examined. The proportion of CEHD sampling results recorded into IMIS was 38% (51% for detected and 28% for ND measurements). In the models, the detected sampling results were associated with a higher probability of recording into IMIS than ND sampling results, and this difference was similar for panel versus non-panel samples. Probability of recording remained constant from 1984 to 2009 for sampling results measured on

panels but increased for sampling results of single determinations of an agent. Some OSHA regions had probability of recording two times higher than others. No other variables that we examined were associated with a CEHD sampling result being reported in IMIS.

**Conclusions:** Our results indicate that the under-reporting of sampling results in IMIS is differential: ND results (especially those determined from the panels) seem less likely to be recorded in IMIS than other results. It is important to consider both IMIS and CEHD data in order to reduce bias in evaluation of exposures in workplaces inspected by OSHA.

**Keywords:** CEHD; databank; IMIS; OSHA; occupational exposure; statistical exposure model.

## Introduction

Multi-industry occupational exposure databanks are potential sources of historical individual exposure measurements useful for exposure surveillance (Gómez, 1997; Ruttenber *et al.*, 2001; LaMontagne *et al.*, 2002), epidemiological research (Friesen *et al.*, 2012; Peters *et al.*, 2012; Fritschi *et al.*, 2015; Taeger *et al.*, 2015), and as a basis for exposure prediction models (Gabriel, 2006; Scarselli *et al.*, 2007; van Tongeren *et al.*, 2011; Mater *et al.*, 2016). Set up in several countries in the early 1980s, these databanks contain large quantities of exposure data generated by governmental agencies during various regulatory and prevention activities.

In the United States, the Occupational Safety and Health Administration (OSHA) maintains two separate databanks that include measurement results collected during compliance inspections. The Integrated Management Information System (IMIS), recently replaced by the OSHA Information System (OIS) (US Department of Labor, 2014a), contains exposure results from surveys performed by OSHA officers. IMIS has been used to evaluate occupational exposures to various chemical agents (Hamm and Burstyn, 2011; Henn *et al.*, 2011; Cowan *et al.*, 2015; Lee *et al.*, 2015). The second databank, referred to as the Chemical Exposure Health Data (CEHD), was made available in 2010 and contains the analytical sample results of the measurements collected by OSHA officers. Officers interpret the CEHD results and record their assessment in IMIS (e.g. calculating an 8-h time-weighted average (TWA) concentration, see Fig. 1).

Even if IMIS and CEHD represent a great potential source of information, results stored within these databanks cannot be regarded, by default, as representative of the exposures experienced by the general U.S. working population. The process by which OSHA selects workplaces for enforcement visits and workers for exposure monitoring is non-random and may over-represent situations with higher- or lower-than-average exposures (Lavoue *et al.*, 2013; Sarazin *et al.*, 2016).

The availability of the CEHD databank provides the opportunity to ask a new question: is there a difference between the population of situations sampled by OSHA officers (i.e. appearing in the CEHD databank) and the population of results recorded in IMIS?

Two OSHA reports published in the 1980s (Mendeloff, 1984; Jones *et al.*, 1986) commented that not all measurements made by OSHA officers resulting in laboratory samples in CEHD were recorded in IMIS; these findings were recently corroborated (Lavoue *et al.*, 2013; Cowan *et al.*, 2015; Lee *et al.*, 2015). Moreover, Lavoue *et al.* (2013) showed that the proportion of non-detects (ND) in the CEHD dataset was higher than in the IMIS dataset for lead.

The main objective of this study was to explore under-reporting in IMIS comprehensively by identifying its determinants based on the linkage and comparison of CEHD and IMIS across a broad range of chemicals.

## Methods

### The OSHA databanks: IMIS and CEHD

The IMIS exposure databank contains exposure results from all chemical and physical hazards collected under federal OSHA and state OSHA plans (OSHA, 2015b) and information about the company inspected. The chemical exposure results collected under enforcement programs (excluding non-compliance sampling such as health consultations and disasters) were accessed through a Freedom of Information Act request. Date, sample number, sample type, and type of inspection are also recorded.

The CEHD databank was accessed online (OSHA, 2015a). Additional information contained in CEHD but not recorded in the IMIS databank includes the following: sampling duration, analytical method, and presence of other substances on the same sampling media. Because data in CEHD are mainly federal OSHA data, there are measurements in IMIS from State OSHA plans that are not in CEHD.

### Data preparation

The IMIS and CEHD datasets were restricted to overlapping years of 1984–2009. The IMIS extract contained

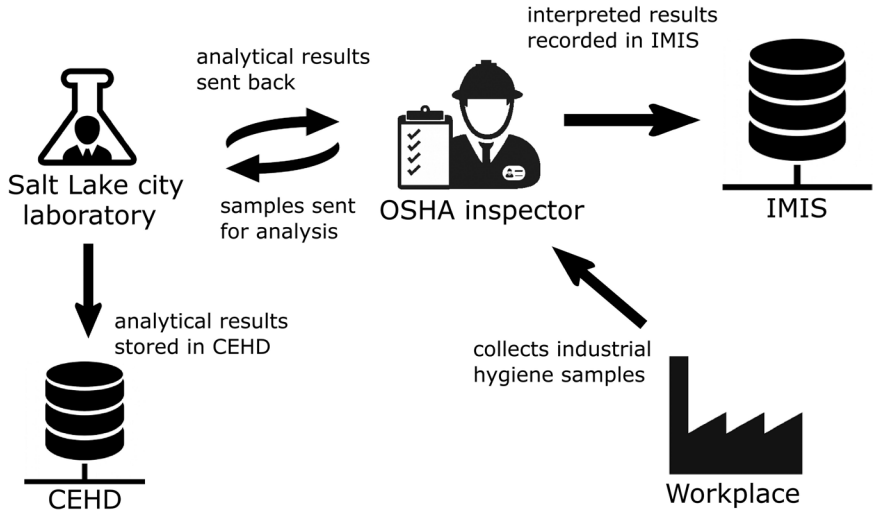


**Figure 1.** Conceptual linkage between the CEHD and IMIS databanks.

851 987 records corresponding to 107 647 inspections, covering 1054 agents. Cleaning of IMIS data was described in Sarazin *et al.* (2016). The CEHD online dataset contained 1 908 373 records corresponding to 40 158 inspections, covering 1082 agents. Cleaning of CEHD data was described in detail in appendix 1 in Lavoue *et al.* (2013). Briefly, records that were judged not useful for exposure assessment (e.g. 'soil', 'gravimetric determination', 'sample weight' measurements) or erroneous (e.g. records with sampling volume or sampling duration missing or null, records with missing or '0' values for sampling number) were excluded. Blank samples and records that were not personal samples (e.g. area, wipe, and bulk) were also excluded. The analyses were restricted to chemical agents that had at least 2500 CEHD sampling results (these agents represent 82% of all CEHD sampling results).

## Linkage between IMIS and CEHD

The 'sampling number' and 'chemical agent' variables were present in both datasets, and a unique 'sampling number–chemical agent' identifier was created to link the two datasets. Sampling number is a unique identifier for the Air Sampling Worksheet (OSHA 91A), which contains inspector's sampling field notes and serves as a submission document for samples requiring analysis at the Salt Lake City Laboratory. This identifier usually determines a unique 'evaluation' made by an inspector for a worker (e.g. a worker's full shift). Several samples in CEHD for one 'sampling number–chemical agent' would correspond to partial-shift measurements (e.g. morning and afternoon samples) aggregated by the inspector before recording in IMIS through the calculation of a TWA.

This identifier was not perfect for linking CEHD to IMIS. For instance, in some cases, aggregated sampling time seemed unrealistically high (>600 min), suggesting several workers were monitored. Moreover, some records in IMIS have the same 'sampling number-agent', corresponding to several sample types [i.e. one shift-long TWA result and one short-term (ST) result]. Hence some CEHD records tied to one 'sampling number–agent' might be used to calculate a TWA result in IMIS, whereas the remaining CEHD record is used to calculate a ST result in IMIS (there is no formal link to identify which CEHD records are associated to which sample type in IMIS, and duration in CEHD is not useful when the IMIS sample type value is 'ND'). However, these issues affected a small proportion of the data, with 7% of the CEHD aggregated results with duration > 8 h and 3% of the IMIS records associated with more than one sample type per sampling number.

Multiple records tied to a single 'sampling number–chemical agent' in CEHD were therefore treated as sequential partial-shift measurements and aggregated to calculate total sampling time and a TWA concentration result for the evaluation. When one of the samples was reported as a ND, its value was replaced by 0 in the calculation of the TWA concentration. If all samples were ND, the aggregated value was reported as a ND.

For CEHD, the terms 'measurements' and 'sampling results' were used to designate the individual analytical records and aggregated TWA concentration results, respectively.

### Ancillary information examined

The focus of this study was to identify the factors associated with a CEHD sampling result being present or not in IMIS. The primary interest was to investigate whether this was related to the measured exposure level (e.g. are ND records less likely to be present in IMIS? Or are high exposure levels more likely to be present in IMIS?).

For each chemical agent, the CEHD sampling results were divided into three categories indicating whether the results were detected and whether they exceeded the permissible exposure limit (PEL) of the agent at the time of measurement: ND, detected below PEL (detected < PEL), and equal or above PEL (detected ≥ PEL). The ST PEL was used to select the appropriate category for sampling results with a total sampling time of ≤15 min for the six agents with both TWA and ST PEL limits (beryllium, cadmium dust, cadmium fume, chromic acid, styrene, toluene) (OSHA, 2014).

Several reports have indicated that multiple agents, such as metals, are often measured on the same sample media as part of a panel (Okun *et al.*, 2004; Hamm and Burstyn, 2011; Henn *et al.*, 2011; Lavoue *et al.*, 2013). In such a case, it would be difficult to know whether a ND result reflected a sample where the agent of interest was not detected or a sample where the agent was not investigated but analyzed nevertheless. We suspected that ND measurements within a panel were only reported by the laboratory because of analytical protocol and would therefore not tend to be recorded into IMIS. The variable 'field number' in the CEHD databank identifies measurements collected on the same sampling media. CEHD sampling results were therefore divided into two categories indicating whether the result belonged to a panel (panel = yes if more than one agent on the sampling media).

We also suspected that CEHD sampling results might be more likely to be recorded into IMIS when other samples taken during the same inspection have detected results. The 'other detected samples in inspection' variable was derived by looking at the list of sampling results in each CEHD inspection and calculating the proportion that were detected. This variable was analyzed as a four-level categorical variable (0% = none, 1–33% = low proportion, 34–66% = medium proportion, more than ≥66% = high proportion).

The number of workers in each inspected facility was categorized into tertiles observed in the CEHD dataset: 1–35 workers = small, 36–150 = medium, ≥150 = large. The sampling duration variable was standardized by subtracting the mean from the value of each record and dividing by the standard deviation.

We used the publicly available IMIS violation dataset (US Department of Labor, 2014b) to create four variables associated with the violative behaviour of a given establishment. The variable 'PEL citations' represents the number of citations issued in the following categories during the inspection: Occupational Safety and Health (OSH) Standards from 1910.1000 through 1910.1052, OSH Standards for Shipyard Employment from 1915.1000 to 1915.1050, and OSH Standards for Construction from 1926.1101 to 1926.1148. The variable 'respiratory protection and hazard communication citations' represents citations related to OSH Standards 1910.0134 and 1910.1200. The third variable created ('other citations') included all other citations (mainly associated with safety issues and physical hazards). A list of OSH Standards codes along with a short description for each is available within the online supplementary material (see Supplementary Table S1, available at the *Annals of Work Exposures and Health* online). Each of the three previous variables was analyzed as a three-level categorical variable ('no' category plus two categories obtained from separating the non-zero values by their median). Finally, we created the 'penalty' variable, representing the total amount of fines historically received by an establishment (the initial penalty fines were used to calculate the amount). This variable was analyzed as a four-level categorical variable ('no penalty' category plus three categories obtained from separating the non-zero values by their tertiles).

Before modeling, correlations between independent variables were evaluated using Cramer's V (Fisher and van Belle, 1993), with a threshold of 0.7 for detecting potential multicollinearity.

### Statistical modeling

We used Poisson regression using a sandwich variance estimator for estimation of risk ratios (RRs) with binary outcome (Greenland, 2004; Zou, 2004) to model the probability of a CEHD sampling result being recorded into IMIS for each individual chemical agent. All variables described above were added in the models, as well as sampling year and OSHA region (Table 1). In addition, we added the interaction of exposure level with panel sample to specifically test whether ND results part of a panel are recorded differently into IMIS, i.e. results coming out of the laboratory, but probably not of interest to the OSHA officer. For OSHA region, we used deviation coding (Menard, 2002; UCLA, 2015) to allow comparisons of the probability of a CEHD sampling result being recorded into IMIS for a given level of the variable to the overall mean probability of recording of the variable [e.g. comparing level 1 (Boston) of OSHA region to the average across all levels of OSHA region, comparing level 2 (New York) of OSHA region to the average across

**Table 1.** IMIS and CEHD ancillary variables tested in the empirical statistical models.

| Variable | Description | Type | Number of samples (%) |
|---|---|---|---|
| Exposure level | Level of exposure of CEHD sampling result | Nominal (three categories) | |
| | | (1)ND | 266 607 (58) |
| | | (2)Detected < PEL[a] | 172 862 (37) |
| | | (3)Detected ≥ PEL | 22 431 (5) |
| Panel sample | CEHD sampling result is part or not of a panel of samples | Nominal (two categories) | |
| | | (1)No | 62 274 (13) |
| | | (2)Yes | 399 626 (87) |
| Year | Year of sampling | Continuous (integer) | |
| | | 1984–2009 | |
| Sampling time | Duration of sampling of CEHD sampling result in minutes | Continuous (integer) | |
| | | Interquartile range = [262;450] | |
| PEL citations | Number of PEL citations issued during the inspection | Nominal (three categories) | |
| | | (1)None | 211 991 (46) |
| | | (2)Low (1–4 citations) | 140 640 (30) |
| | | (3)High (5+ citations) | 109 269 (24) |
| Respiratory protection and hazard communication citations | Number of respiratory protection and hazard communication citations issued during the inspection | Nominal (three categories) | |
| | | (1)None | 171 354 (37) |
| | | (2)Low (1–3 citations) | 172 687 (37) |
| | | (3)High (4+ citations) | 117 859 (26) |
| Other citations | Number of other types of citations issued during the inspection | Nominal (three categories) | |
| | | (1)None | 102 746 (22) |
| | | (2)Low (1–5 citations) | 205 123 (44) |
| | | (3)High (6+ citations) | 154 031 (33) |
| Detected samples in inspection | Proportion of detected sampling results in the inspection | Nominal (four categories) | |
| | | (1)None | 19 559 (4) |
| | | (2)Low (1–33%) | 191 857 (42) |
| | | (3)Med (34–67%) | 206 946 (45) |
| | | (4)High (68–100%) | 43 538 (9) |
| Establishment size | Number of employees working in the establishment | Nominal (three categories) | |
| | | (1)Small (1–35 employees) | 152 189 (33) |
| | | (2)Medium (36–150 employees) | 165 341 (36) |
| | | (3)Large (151+ employees) | 144 370 (31) |
| Penalty | Sum of historical penalties assessed in the establishment monitored | Nominal (four categories) | |
| | | (1)None | 35 717 (8) |
| | | (2)Low | 139 321 (30) |
| | | (3)Medium | 142 039 (31) |
| | | (4)High | 144 823 (31) |
| OSHA region[b] | Identifies the OSHA region where the inspection took place | Nominal (10 categories) | |
| | | (1)01_boston | 38 166 (8) |
| | | (2)02_new_york | 57 810 (13) |
| | | (3)03_philadelphia | 41 059 (9) |
| | | (4)04_atlanta | 51 144 (11) |
| | | (5)05_chicago | 168 022 (36) |
| | | (6)06_dallas | 50 429 (11) |
| | | (7)07_kansas_city | 16 838 (4) |
| | | (8)08_denver | 25 592 (6) |
| | | (9)09_san_francisco | 9623 (2) |
| | | (10)10_seattle | 3217 (1) |

[a]PEL at time of measurement.

[b]https://www.osha.gov/html/RAmap.html.

all levels of OSHA region, and so on]. The curvilinear relationship of year with probability of recording was tested using polynomial functions. The complexity of the polynomial was determined by increasing the degrees of the polynomial from 1 to 4 until no additional improvement in model fit was observed [based on the Akaike Information Criterion (Burnham and Anderson, 2002)]. Meta-analytic methods were used as described in Sarazin *et al.* (2016) to combine results from all chemical agents (van Houwelingen *et al.*, 2002; Borenstein *et al.*, 2010).

To assess whether the effect of the level of exposure and panel status on the probability of a CEHD sampling result being recorded into IMIS changed across time, an additional Poisson regression model was fitted to the full dataset with the same structure as the agent-by-agent analysis, but with an interaction of year with exposure level and panel status (this model structure could not be applied to individual agent datasets because of low sample size). This model approach implied different recording probabilities for each agent but assumed that the magnitude of the influence of other predictors was the same across agents.

### Software

All analyses were performed using the R 3.1.3 statistical software (R Development Core Team, Vienna, Austria), with the package *metafor* (Viechtbauer, 2014) for the meta-analysis, *ggplot2* (Wickham, 2015) for graphical illustrations, *vcd* (Meyer, 2015) for computing of Cramer's V coefficients, and *sandwich* (Zeileis, 2015) for calculation of robust standard error estimators.

## Results

### Descriptive analysis

The analyses included 728 127 CEHD analytical measurements corresponding to 28 179 inspection visits for the period 1984 to 2009. 297 868 measurements had a 1:1 link with a sampling number (sampling time range: 8–568 min). The remaining 430 259 measurements were aggregated to calculate 164 032 sampling results (sampling time range: 19–850 min), yielding a total of 461 900 CEHD sampling results.

Twenty-four agents met our inclusion criteria for analysis, constituting 82% of all CEHD personal sampling results (14 metals and their compounds, 5 organic solvents, 3 dusts/fibers, and 2 other agents) (Table 2). The proportion of aggregated sampling results was respectively 33% and 72% for metals and solvents. Among aggregated results, metals and dusts had a median of 2 samples per aggregated result, whereas solvents had a median of 4. The overall proportion of CEHD sampling results recorded into IMIS was 38% (51% for detected records and 28% for ND records). Lead, 4,4'-methylene diphenyl diisocyanate (MDI), and Stoddard solvent had the highest proportion of sampling results recorded into IMIS (65%, 59%, 54%, respectively), whereas cadmium dust and particulates not otherwise regulated (PNOR) – respirable dust had less than 10% of their sampling results recorded. Crude rates of sampling results recorded into IMIS for all variables examined in this study are available within the online supplementary material (see Supplementary Table S2, available at the *Annals of Work Exposures and Health* online).

### Statistical modeling

Most independent variable pairs had weak correlation based on Cramer's $V$ ($r < 0.4$), except for the panel sample/chemical agent and exposure level/chemical agent pairs ($r = 0.66$ and $r = 0.48$, respectively). The association of year with probability of a CEHD sampling result being recorded into IMIS was best modeled with fourth degree polynomial. Increasing the degrees of the polynomial from 1 to 4 resulted in a reduction of the Akaike information criterion (AIC) for all agents (1 degree = reference, 2 degrees = median of −400 AIC units across agents, 3 degrees = median of −600 AIC units across agents, and 4 degrees = median of −1480 AIC units across agents).

Table 3 shows the observed pooled association of all categorical predictor variables with the probability of recording a CEHD sampling result in IMIS as meta-analytic RRs. The detected sampling results were associated with a higher probability of recording into IMIS than ND sampling results, and this difference was similar for panel versus non-panel samples. Results that were equal to or above PEL had a slightly higher probability of being recorded into IMIS than detected levels below the PEL. The ND sampling results measured as part of a panel were associated with a lower probability of recording into IMIS than ND sampling results not part of a panel. Probability of recording varied between OSHA regions by ±40% relative to the national average.

Forest plots useful to evaluate how agent-specific associations relate to the pooled estimate for each level of each predictor variable are available within the online supplementary material (see Supplementary Figure S1, available at the *Annals of Work Exposures and Health* online). As an example, we show in Fig. 2 the agent-specific and meta-analytic RRs for sampling results that were equal or greater than PEL compared with ND sampling results stratified by panel status. Visual assessment of forest plots generally showed homogeneity across agents for all predictors, with a few exceptions of

**Table 2.** Descriptive statistics of chemicals in CEHD and IMIS meeting the inclusion criteria.

| Chemical agent | CEHD | | | | | IMIS |
|---|---|---|---|---|---|---|
| | Number of sampling results | Proportion of aggregated sampling results [median number of measurements per aggregated sampling result][a] | ND (%) | Proportion of sampling results ≥ PEL (%) | Proportion of sampling results recorded into IMIS (%) | Number of exposure results |
| Organic solvents | | | | | | |
| Toluene | 9030 | 72 [3] | 10 | 1 | 53 | 22 063 |
| Xylene | 7987 | 69 [3] | 15 | 2 | 52 | 14 366 |
| 2-butanone | 3355 | 81 [4] | 17 | 3 | 49 | 6892 |
| Stoddard solvent | 2734 | 70 [4] | 33 | 0 | 54 | 3629 |
| Acetone | 2567 | 74 [4] | 15 | 1 | 53 | 6086 |
| Metals and their compounds | | | | | | |
| Lead, inorganic | 48 291 | 33 [2] | 58 | 20 | 65 | 59 700 |
| Copper fume | 33 663 | 33 [2] | 36 | 7 | 42 | 25 697 |
| Zinc oxide fume | 33 212 | 33 [2] | 36 | 1 | 40 | 25 023 |
| Antimony | 33 137 | 33 [2] | 95 | 0 | 29 | 13 612 |
| Chromium | 33 035 | 33 [2] | 61 | 1 | 36 | 20 208 |
| Beryllium | 32 783 | 33 [2] | 96 | 1 | 29 | 14 478 |
| Nickel | 32 641 | 33 [2] | 75 | 1 | 33 | 18 380 |
| Iron oxide fume | 32 492 | 33 [2] | 15 | 4 | 47 | 32 145 |
| Cobalt | 32 410 | 33 [2] | 89 | 1 | 31 | 14 985 |
| Manganese fume | 32 283 | 33 [2] | 36 | 0 | 40 | 24 716 |
| Cadmium fume | 15 869 | 36 [2] | 88 | 1 | 33 | 10 369 |
| Cadmium dust | 14 193 | 33 [2] | 81 | 5 | 5 | 3334 |
| Arsenic | 4364 | 39 [2] | 52 | 11 | 27 | 3004 |
| Chromic acid | 2834 | 32 [2] | 52 | 7 | 48 | 3928 |
| Dust/fibers | | | | | | |
| PNOR (respirable dust) | 23 129 | 33 [2] | 82 | 2 | 5 | 12 092 |
| Silica, quartz | 12 978 | 25 [2] | 27 | 28 | 53 | 28 172 |
| PNOR (total dust) | 12 036 | 29 [2] | 28 | 8 | 23 | 18 980 |
| Other | | | | | | |
| Styrene | 3819 | 78 [4] | 9 | 11 | 53 | 9152 |
| 4,4'-MDI | 3058 | 53 [2] | 65 | 12 | 59 | 4360 |

[a]Percentage of aggregated sampling results and median number of individual CEHD measurements per aggregated TWA sampling result.

**Table 3.** Summary of meta-analytic RRs for the probability of a CEHD sampling result being recorded into IMIS.

| Variable/category | RR (95% CI) |
|---|---|
| (Panel) × (exposure level) | |
| Panel_no: ND | 1.00 (reference)[a] |
| Panel_no: Detected < PEL | 1.35 (1.23;1.47)[b] |
| Panel_no: Detected ≥ PEL | 1.51 (1.33;1.71) |
| Panel_yes: ND | 0.88 (0.80;0.96) |
| Panel_yes: Detected < PEL | 1.25 (1.14;1.37) |
| Panel_yes: Detected ≥ PEL | 1.36 (1.20;1.56) |
| Sampling time | |
| 480 min versus 30 min | 1.12 (1.09;1.14)[c] |
| PEL citations in inspection | |
| None (0) | 1.00 (reference) |
| Low (1–4) | 0.99 (0.97;1.01) |
| High (5+) | 0.96 (0.93;0.99) |
| RespProt and HazComm citations in inspection | |
| None (0) | 1.00 (reference) |
| Low (1–3) | 1.03 (1.01;1.05) |
| High (4+) | 1.05 (1.02;1.08) |
| Other citations in inspection | |
| None (0) | 1.00 (reference) |
| Low (1–5) | 1.04 (1.01;1.07) |
| High (6+) | 1.01 (0.99;1.04) |
| Other detected level in inspection | |
| None (0%) | 1.00 (reference) |
| Low (1–33%) | 1.01 (0.97;1.06) |
| Med (34–66%) | 0.99 (0.97;1.01) |
| High (67%+) | 0.97 (0.93;1.00) |
| Establishment size | |
| Small (1–35) | 1.00 (reference) |
| Medium (36–150) | 1.00 (0.99;1.02) |
| Large (151+) | 0.95 (0.94;0.97) |
| Penalty | |
| None | 1.00 (reference) |
| Low | 1.01 (0.98;1.03) |
| Medium | 0.99 (0.96;1.02) |
| High | 0.96 (0.92;1.00) |
| OSHA region[d] | |
| Mean of OSHA regions | 1.00 (reference) |
| Boston | 1.07 (1.04;1.09) |
| New York | 0.74 (0.69;0.79) |
| Philadelphia | 0.78 (0.71;0.85) |
| Atlanta | 0.94 (0.90;0.99) |
| Chicago | 1.00 (0.98;1.03) |
| Dallas | 1.13 (1.11;1.16) |
| Kansas City | 1.16 (1.11;1.21) |
| Denver | 1.39 (1.30;1.49) |
| San Francisco | 0.86 (0.83;0.89) |
| Seattle | 1.21 (1.06;1.37) |

[a]RR of the reference levels taken as 1.
[b]95% CI.
[c]RR for sampling time 480 min compared with 30 min.
[d]https://www.osha.gov/html/RAmap.html.

note: issuance of PEL citations during an inspection was associated with higher probability of recording for cadmium dust (agent-specific RR = 1.63, 95% confidence interval (CI): 1.33–2.01; meta-analytic RR = 0.99, 95% CI: 0.97–1.01), whereas a higher proportion of detected sampling results in the inspection was associated with higher probability of recording for cadmium fume (agent-specific RR = 1.52, 95% CI: 1.15–2.01; meta-analytic RR = 0.97, 95% CI: 0.93–1.00). A high total amount of penalty was associated with higher probability of recording for acetone (agent-specific RR = 1.56, 95% CI: 1.31–1.86; meta-analytic RR = 0.96, 95% CI: 0.92–1.00).

Similar to the agent-by-agent analysis, the full dataset approach showed that exposure level, panel sample, and OSHA region were the most strongly associated with probability of recording (see Supplementary Table S3, available at the *Annals of Work Exposures and Health* online). Supplementary Figure S2 (available at the *Annals of Work Exposures and Health* online) shows the predicted probability of recording for year stratified by exposure level and panel sample status. Visual assessment of the smoothed curves suggests greater increase from 1984 to 2009 for detected (RR increase: ~0.6) compared with ND sampling results (RR increase: ~0.2) when measured alone. The probability of recording remained fairly constant from 1984 to 2009 for detected sampling results measured on panels, whereas a small decrease in probability was seen for ND sampling results measured on panels (RR decrease: ~0.1) in more recent years. Across the 24 agents, the probability of a CEHD sampling result being recorded into IMIS varied between 0.16 (PNOR – respirable dust) and 2.00 (lead) compared with the overall average, with an interquartile range = 1.12; 1.26 (see Supplementary Table S4, available at the *Annals of Work Exposures and Health* online).

## Discussion

We expanded on the initial analyses performed by Lavoue *et al.* (2013), Jones *et al.* (1986), and Mendeloff (1984) by looking at a broad range of chemical agents (82% of the CEHD databank included) and using statistical modeling to study concomitantly several potential explanatory variables.

The overall proportion of CEHD sampling results recorded into IMIS was 38% for the period 1984–2009, with a higher proportion of recording for detected results (51%) compared with ND results (28%). The results from the multivariate regression models showed that level of exposure, panel sample status, year of sampling, and

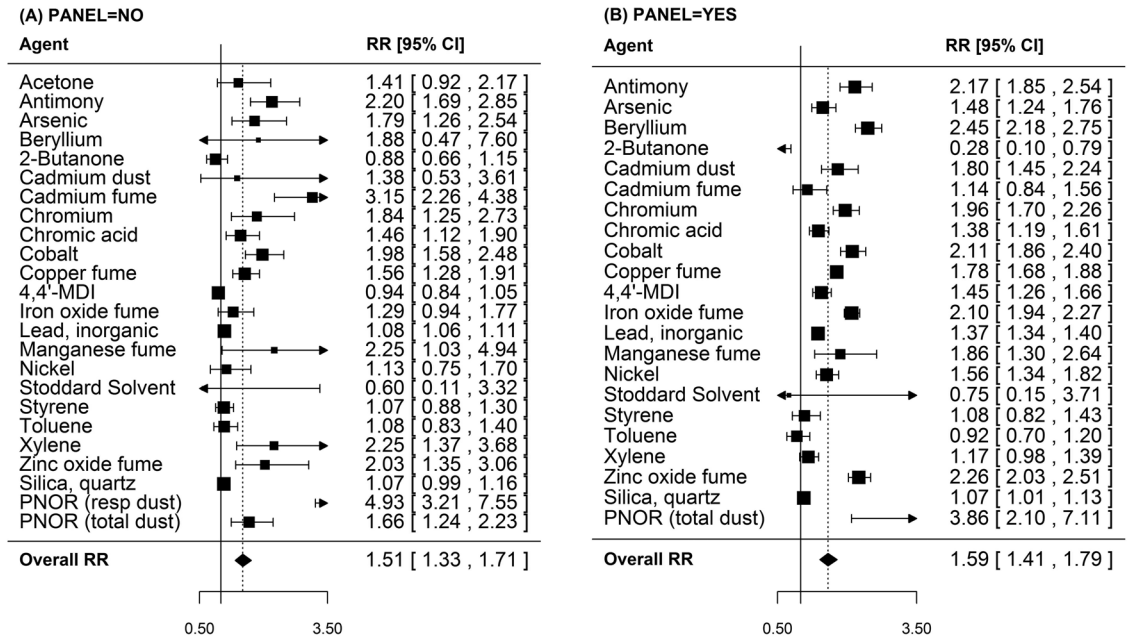


**Figure 2.** Agent-specific and meta-analytic RRs for the probability of a CEHD sampling result being recorded into IMIS for 'exposure level ≥ PEL' compared with 'exposure level = ND'. Agent-specific RRs were pooled with the random-effects method. Squares represent agent-specific risk estimates (size of the square reflects the agent-specific statistical weight); horizontal lines, the 95% CIs; diamond, the summary risk estimate and its corresponding 95% CI. For (A), the reference level is panel_no: ND; for (B), the reference level is panel_yes: ND.

OSHA region were the variables most strongly related to the CEHD sampling result being recorded into IMIS.

Higher probability of recording of detected versus ND sampling result was seen regardless if measured on panels or alone. Our findings suggest that ND sampling results might be considered by an OSHA officer as 'not worth' reporting in IMIS. Moreover, ND sampling results from chemicals measured on panels were even less frequently recorded, which is plausible since they would only be analyzed because of analytical protocol and not of *a priori* interest for risk analysis. On the other hand, the probability of recording was generally the same for detected sampling results regardless if measured alone or on a panel, which is consistent with the fact that such results likely reflect an agent being investigated by the officer. Our observations also point to a slightly higher probability of exposure levels ≥PEL being recorded in IMIS compared with detected levels <PEL, likely reflecting the importance to report higher exposure levels by OSHA officers. The overall signal is however relatively weak compared with the difference ND/detected, and we could not distinguish patterns or groups of agents with a clearly stronger association. Our results support the early hypotheses of Jones *et al.* (1986) and Mendeloff (1984) that the IMIS under-reporting is differential.

Sampling results with higher sampling time were more likely to be recorded into IMIS. These observations might be explained by the fact that OSHA inspectors are more likely to record a result in IMIS that is deemed to be more informative for assessment of compliance.

The differences observed between regions might be explained by varying practices of compliance officers. We did explore whether industry might explain some of these differences by fitting our models (results not shown) with an added broad industry classification but did not observe any influence. Finer analyses performed on single agents might help shed more light on this issue.

It was expected that sampling results obtained during inspections for which citations were issued might be more systematically recorded, but no consistent trend was observed. Recording was lower for PEL citations, higher for respiratory protection/other types of citations, and lower for historical amount of penalty assessed to the establishment, but these RRs were close to unity.

The associations between the ancillary variables and a CEHD sampling result being reported in IMIS found in this study might not be applicable to the agents that were not analyzed, but nevertheless reflect the overwhelming majority of inspections performed by OSHA inspectors and measurements collected by them over the years.

Similar to the agent-by-agent analysis, the results from the additional full dataset analysis showed that level of exposure, panel sample status, year of sampling, and OSHA region were the variables most strongly related to the CEHD sampling result being recorded into IMIS. Analysis of time trends showed an increase over time in the proportion of recording for sampling results measured alone (detected and ND), flat slope for detected sampling results measured as part of panels, and a small decrease for ND sampling results measured as part of panels. Moreover, there were differences beyond chance between chemical agents in probability of recording. They likely correspond to specific programs that emphasized certain agents since the setup of OSHA databanks in the 1970s (e.g. highest probability of recording for lead) or to particular sampling technique circumstances (e.g. very low probability of recording for dust samples). It is probable that most dust measurements in the CEHD databank are only present because the officer requested another agent. For example, total dust is needed in order to get the measurement of a metal, but the officer would not report total dust in IMIS since it provides no information about the assessment of interest – metals. This hypothesis is supported by the fact that the crude proportion of recording was 35% for total dust sampling results measured alone compared with 9% when measured on a panel (usually with metals). Fitting the model to the full dataset restricted to the 77 agents with >1000 instead of 2500 records yielded the same overall results.

It is difficult to evaluate whether under-recording occurs for other occupational databases, and we are not aware of other papers addressing this issue. For illustration purpose, in the French COLCHIC databank, linkage between the laboratory results and the databank itself is automated (Vincent and Jeandel, 2001).

### Limitations

Some limitations need to be acknowledged. First, the lack of information on the unit monitored (e.g. worker, task, and tool) may have misled the aggregation procedure used to regroup sequential CEHD measurements. Sampling number was used as a single identifier of what we called an 'evaluation', but CEHD measurements linked to one sampling number could correspond to the monitoring of different workers or tasks and would have had to be split into several sample types in IMIS in some cases. However, 93% of the aggregated sampling durations were below the standard 480 min working shift. Despite these reassuring overall results it is probable that for some specific agents often evaluated for both TWA and ST exposures (toluene and styrene), we mistakenly aggregated some long-term and ST results together. However, this would only potentially affect, for a minority of agents, our coarse exposure level variable and the estimation of the effect of the duration variable on recording (visual assessment of forest plots did not show a different behaviour for toluene and styrene compared with other agents in our analyses). We however would recommend development, for agents such as styrene, of a finer linkage approach in analyses focused on estimation of exposure levels.

Second, the interpretation of results is limited by the lack of information on specific practices regarding how data are reported into IMIS. Although we have empirical evidence, the lack of information on region-specific practices (e.g. unwieldiness of computer system) or quality of the sampling data (e.g. judgment of the inspector) prevented an in-depth understanding of the reasons for reporting a sampling result in IMIS.

Third, the conclusions found in this study are applicable to exposure data collected by the federal OSHA inspectors and processed by the Salt Lake Technical Center. Collecting datasets containing laboratory analyses performed at OSHA's state laboratories may be helpful in determining whether under-reporting also exists with state datasets, as some evidence of such phenomenon was reported in the 2009 Federal Annual Monitoring and Evaluation (FAME) report for the state of Indiana (OSHA, 2016).

## Conclusions

The under-reporting of ND sampling results found in this study suggests a deficit in the number of extremely low values in IMIS, which would contribute to an upward bias of exposure levels in the databank. On the other hand, findings also suggest that the level of detected results and the issuance of a citation for overexposure were not involved in the decision to record a sampling result in IMIS. Our analyses indicate that it is important to combine IMIS and CEHD when assessing occupational exposures. Future analyses involving only OIS data (OIS started in 2011 in some regions) will be facilitated as OIS is a web-based platform that is linked to the analytical laboratory.

## Supplementary Data

Supplementary data are available at *Annals of Work Exposures and Health* online.

## Acknowledgements

The authors would like to acknowledge Dan Vatnik for assisting with linking IMIS exposure dataset with the IMIS violation dataset and the OSHA Directorate of Information Technology for providing the IMIS data. P.S. was supported by the Institut de recherche Robert-Sauvé en santé et en sécurité du travail (IRSST). M.C.F was supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics of the National Cancer Institute. The authors declare that there are no conflicts of interest relating to the material in relation to this article.

## References

Borenstein M, Hedges LV, Higgins JP *et al*. (2010) A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res Synth Methods*; **1**: 97–111.

Burnham KP, Anderson DR. (2002) *Model selection and multimodel inference: a practical information-theoretic approach*. New York, NY: Springer.

Cowan DM, Cheng TJ, Ground M *et al*. (2015) Analysis of workplace compliance measurements of asbestos by the U.S. Occupational Safety and Health Administration (1984-2011). *Regul Toxicol Pharmacol*; **72**: 615–29.

Fisher L, van Belle G. (1993) *Biostatistics: a methodology for the health sciences*. New York, NY: John Wiley and Sons.

Friesen MC, Coble JB, Lu W *et al*. (2012) Combining a job-exposure matrix with exposure measurements to assess occupational exposure to benzene in a population cohort in Shanghai, China. *Ann Occup Hyg*; **56**: 80–91.

Fritschi L, Benke G, Risch HA *et al*. (2015) Occupational exposure to N-nitrosamines and pesticides and risk of pancreatic cancer. *Occup Environ Med*; **72**: 678–83.

Gabriel S. (2006) The BG measurement system for hazardous substances (BGMG) and the exposure database of hazardous substances (MEGA). *Int J Occup Saf Ergon*; **12**: 101–4.

Gómez MR. (1997) Factors associated with exposure in Occupational Safety and Health Administration data. *Am Ind Hyg Assoc J*; **58**: 186–95.

Greenland S. (2004) Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *Am J Epidemiol*; **160**: 301–5.

Hamm MP, Burstyn I. (2011) Estimating occupational beryllium exposure from compliance monitoring data. *Arch Environ Occup Health*; **66**: 75–86.

Henn SA, Sussell AL, Li J *et al*. (2011) Characterization of lead in US workplaces using data from OSHA's Integrated Management Information System. *Am J Ind Med*; **54**: 356–65.

Jones C, Weld L, Gray W *et al*. (1986) *The sampling and reporting processes in OSHA MIS data*. Cincinnati, OH: United States National Institute for Occupational Safety and Health, Grant No R03-OH-002135.

LaMontagne AD, Herrick RF, Van Dyke MV *et al*. (2002) Exposure databases and exposure surveillance: promise and practice. *AIHA J (Fairfax, Va)*; **63**: 205–12.

Lavoue J, Friesen MC, Burstyn I. (2013) Workplace measurements by the US Occupational Safety and Health Administration since 1979: descriptive analysis and potential uses for exposure assessment. *Ann Occup Hyg*; **57**: 77–97.

Lee DG, Lavoué J, Spinelli JJ *et al*. (2015) Statistical modeling of occupational exposure to polycyclic aromatic hydrocarbons using OSHA data. *J Occup Environ Hyg*; **12**: 729–42.

Mater G, Paris C, Lavoué J. (2016) Descriptive analysis and comparison of two French occupational exposure databases: COLCHIC and SCOLA. *Am J Ind Med*; **59**: 379–91.

Menard S. (2002) *Applied logistic regression analysis*. Thousand Oaks, CA: SAGE publication.

Mendeloff J. (1984) *A new strategy for estimating occupational exposures to toxic substances*. Cincinnati, OH: United States National Institute for Occupational Safety and Health (microfiche number NIOSH-00182240).

Meyer D. (2015) Visualizing categorical data. Available at https://cran.r-project.org/web/packages/vcd/vcd.pdf. Accessed 15 September 2015.

Okun A, Cooper G, Bailer AJ *et al*. (2004) Trends in occupational lead exposure since the 1978 OSHA lead standard. *Am J Ind Med*; **45**: 558–72.

OSHA. (2014) Permissible exposure limits – annotated tables. Available at https://www.osha.gov/dsg/annotated-pels/. Accessed 2 November 2014.

OSHA. (2015a) Chemical exposure health data. Available at https://www.osha.gov/opengov/healthsamples.html. Accessed 2 August 2015.

OSHA. (2015b) State plans – office of state programs. Available at https://www.osha.gov/dcsp/osp/. Accessed 2 August 2015.

OSHA. (2016) Federal Annual Monitoring and Evaluation (FAME) Reports/Indiana FAME reports. Available at https://www.osha.gov/dcsp/osp/efame/indiana.html#!2009. Accessed 15 June 2017.

Peters S, Vermeulen R, Olsson A *et al*. (2012) Development of an exposure measurement database on five lung carcinogens (ExpoSYN) for quantitative retrospective occupational exposure assessment. *Ann Occup Hyg*; **56**: 70–9.

Ruttenber AJ, McCrea JS, Wade TD *et al*. (2001) Integrating workplace exposure databases for occupational medicine services and epidemiologic studies at a former nuclear weapons facility. *Appl Occup Environ Hyg*; **16**: 192–200.

Sarazin P, Burstyn I, Kincl L *et al*. (2016) Trends in OSHA Compliance Monitoring Data 1979–2011: statistical modeling of ancillary information across 77 chemicals. *Ann Occup Hyg*; **60**: 432–52.

Scarselli A, Montaruli C, Marinaccio A. (2007) The Italian information system on occupational exposure to carcinogens (SIREP): structure, contents and future perspectives. *Ann Occup Hyg*; **51**: 471–8.

Taeger D, Pesch B, Kendzia B *et al*. (2015) Lung cancer among coal miners, ore miners and quarrymen: smoking-adjusted

risk estimates from the synergy pooled analysis of case-control studies. *Scand J Work Environ Health*; **41**: 467–77.

UCLA. (2015) R Library: contrast coding systems for categorical variables. Available at http://www.ats.ucla.edu/stat/r/library/contrast_coding.htm#DEVIATION. Accessed 15 January 2016.

US Department of Labor. (2014a) OSHA Information System (OIS). Available at http://www.dol.gov/oasam/ocio/programs/pia/osha/OSHA-OIS.htm. Accessed 15 October 2015.

US Department of Labor. (2014b) OSHA enforcement data. Available at http://ogesdw.dol.gov/views/data_summary.php. Accessed 15 October 2015.

van Houwelingen HC, Arends LR, Stijnen T. (2002) Advanced methods in meta-analysis: multivariate approach and meta-regression. *Stat Med*; **21**: 589–624.

van Tongeren M, Fransman W, Spankie S *et al*. (2011) Advanced REACH Tool: development and application of the substance emission potential modifying factor. *Ann Occup Hyg*; **55**: 980–8.

Viechtbauer W. (2014) Meta-analysis package for R. Available at http://cran.r-project.org/web/packages/metafor/index.html. Accessed 15 November 2015.

Vincent R, Jeandel B. (2001) COLCHIC-occupational exposure to chemical agents database: current content and development perspectives. *Appl Occup Environ Hyg*; **16**: 115–21.

Wickham H. (2015) Package 'ggplot2'. Available at https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf. Accessed 15 November 2014.

Zeileis A. (2015) Robust covariance matrix estimators. Available at https://cran.r-project.org/web/packages/sandwich/sandwich.pdf. Accessed 15 November 2015.

Zou G. (2004) A modified Poisson regression approach to prospective studies with binary data. *Am J Epidemiol*; **159**: 702–6.