Ann. Occup. Hyg., Vol. 57, No. 1, pp. 77–97, 2013

doi:10.1093/annhyg/mes055

Workplace Measurements by the US Occupational Safety and Health Administration since 1979: Descriptive Analysis and Potential Uses for Exposure Assessment

J. LAVOUE[1], M.C. FRIESEN[2] and I. BURSTYN[3]

[1]University of Montreal Hospital Research Center, Montréal, Québec, Canada; [2]Occupational and Environmental Epidemiology Branch, Division of Cancer Epidemiology & Genetics, National Cancer Institute, North Bethesda, MD; [3]Department of Environmental and Occupational Health, School of Public Health, Drexel University, Philadelphia, PA

Received 29 January 2012; in final form 15 May 2012; published online 5 September

Background: Inspectors from the US Occupational Safety and Health Administration (OSHA) have been collecting industrial hygiene samples since 1972 to verify compliance with Permissible Exposure Limits. Starting in 1979, these measurements were computerized into the Integrated Management Information System (IMIS). In 2010, a dataset of over 1 million personal sample results analysed at OSHA's central laboratory in Salt Lake City [Chemical Exposure Health Data (CEHD)], only partially overlapping the IMIS database, was placed into public domain via the internet. We undertook this study to inform potential users about the relationship between this newly available OSHA data and IMIS and to offer insight about the opportunities and challenges associated with the use of OSHA measurement data for occupational exposure assessment.

Methods: We conducted a literature review of previous uses of IMIS in occupational health research and performed a descriptive analysis of the data recently made available and compared them to the IMIS database for lead, the most frequently sampled agent.

Results: The literature review yielded 29 studies reporting use of IMIS data, but none using the CEHD data. Most studies focused on a single contaminant, with silica and lead being most frequently analysed. Sixteen studies addressed potential bias in IMIS, mostly by examining the association between exposure levels and ancillary information. Although no biases of appreciable magnitude were consistently reported across studies and agents, these assessments may have been obscured by selective under-reporting of non-detectable measurements. The CEHD data comprised 1 450 836 records from 1984 to 2009, not counting analytical blanks and erroneous records. Seventy eight agents with >1000 personal samples yielded 1 037 367 records. Unlike IMIS, which contain administrative information (company size, job description), ancillary information in the CEHD data is mostly analytical. When the IMIS and CEHD measurements of lead were merged, 23 033 (39.2%) records were in common to both IMIS and CEHD datasets, 10 681 (18.2%) records were only in IMIS, and 25 012 (42.6%) records were only in the CEHD database. While IMIS-only records represent data analysed in other laboratories, CEHD-only records suggest partial reporting of sampling results by OSHA inspectors into IMIS. For lead, the percentage of non-detects in the CEHD-only data was 71% compared to 42% and 46% in the both-IMIS-CEHD and IMIS-only datasets, respectively, suggesting differential under-reporting of non-detects in IMIS.

*Author to whom correspondence should be addressed. Email: jerome.lavoue@umontreal.ca

Conclusions: IMIS and the CEHD datasets represent the biggest source of multi-industry exposure data in the USA and should be considered as a valuable source of information for occupational exposure assessment. The lack of empirical data on biases, adequate interpretation of non-detects in OSHA data, complicated by suspected differential under-reporting, remain the principal challenges to the valid estimation of average exposure conditions. We advocate additional comparisons between IMIS and CEHD data and discuss analytical strategies that may play a key role in meeting these challenges.

Keywords: database; exposure reconstruction; IMIS; occupational hygiene.

INTRODUCTION

Recent advances in exposure assessment in occupational epidemiology indicate a shift from approaches based on expert judgement to using objective measurements wherever possible. Industry-based studies focused on a small number of facilities are the best able to incorporate measurements because current and historical exposure data are extracted from a restricted number of sources. For population-based case–control studies, the subjects' occupations span a wide spectrum of activities, representing hundreds, and even thousands, of occupation–industry combinations in a typical study. In this situation, even if measurements had been taken by companies themselves or various governmental agencies over time, the resources needed to collect and interpret such data may be impractically high. In consequence, exposure assessment for population-based studies needs readily available sources of measurements that represent a wide variety of occupations, industries, and time periods to avoid relying solely on expert judgment. Beyond epidemiology, such data can be instrumental for other prevention activities (Gomez, 1993). Potential applications include examining time-trends in exposures (Kromhout and Vermeulen, 2000; Creely *et al*., 2007; Symanski *et al*., 1998), estimating numbers of workers exposed for surveillance efforts or for evaluating the burden of disease caused by an agent (Linch *et al*., 1998; Henneberger *et al*., 2004), identifying high exposure situations to help define intervention priorities (Froines *et al*., 1986), or validating risk assessment tools used to comply with the Registration, Evaluation, Authorization, and Restriction of Chemicals legislation in Europe (Koppisch *et al*., 2012).

Perhaps the greatest potential source of individual measurement data comes from nation-wide occupational exposure databanks. Set up in several countries at the beginning of the 1980s, these databanks contain measurements made by governmental agencies for various purposes including regulatory activities. Countries for which such databanks have been described in the literature include France (Vincent and Jeandel, 2001), United Kingdom (Burns and Beaumont, 1989), Germany (Gabriel, 2006; Koppisch *et al*., 2012), Norway (Lenvik *et al*., 1999), Denmark (Vinzents *et al*., 1995), Finland (Kauppinen, 2001), Singapore (Tang *et al*., 2006), Italy (Scarselli *et al*., 2007), and the United States (Stewart and Rice, 1990). After more than 30 years of data recording for some databanks, the amount of data available has reached a critical mass to permit exposure portraits to be drawn, i.e. to estimate exposure distributions across a wide range of agents, industries, occupations, and years (Kauffer and Vincent, 2007; Lavoué *et al*., 2011, 2008). An alternative data source is the data reported in the published literature, which has been used to support several exposure assessment efforts (e.g. Hein *et al*., 2010; Liu *et al*., 2011; Park *et al*., 2009). However, these data are generally available in aggregate form, require substantial time commitment to extract the data, and have limited ancillary data (Hein *et al*., 2008).

In the USA, the Occupational Safety and Health Administration (OSHA) has maintained since 1979 the Integrated Management Information System (IMIS), which contains measurement results from surveys performed by OSHA to verify compliance to Permissible Exposure Limits (PELs). IMIS, with now over 1.5 million records (Okun *et al*., 2004), is the biggest multi-industry source of exposure measurements in North America. In 2010, OSHA made available on the web[1] all OSHA measurements analysed by the OSHA Salt Lake Technical Center from 1984 through 2009, comprising almost 2 million records (hereafter referred to as the CEHD data, 'Chemical Exposure Health Data'). Although there is overlap in the OSHA measurements contained within the CEHD and IMIS databanks, these two datasets have important differences, which we discuss later.

The IMIS and the CEHD databanks have considerable potential as a source of generic exposure information. Thus, we describe the content of both

[1]http://www.osha.gov/opengov/healthsamples.html

repositories. Our specific aims were to summarize previous uses of IMIS data, to describe the CEHD data and their relationship to the IMIS data, and to highlight strengths and limitations of these databanks. In addition, we make recommendations for the data's future use, with special attention to methodological challenges in occupational exposure assessment.

METHODS

OSHA measurement activities and the IMIS

OSHA was created as a federal agency in 1971 (US Congress, 1970). Some states opted out of the federal OSHA agency and created their own State OSHA agencies, and some states use a combination of federal and State OSHA agencies. Since 1972, IMIS has served as a data-entry and information retrieval system associated with enforcement activities of both federal and State OSHA. Each OSHA inspector is responsible for documenting the outcome of each inspection, including entering exposure measurements into IMIS. The actual exposure levels measured during inspections were only entered starting in 1979. Before that, only a 'severity index' was provided, representing the ratio of the measurement to the PEL. The Salt Lake Technical Center, created in 1984, processed most of the samples collected by the federal and some of the samples collected by State OSHA inspectors. The CEHD data made available by the Salt Lake Technical Center are analytical sample results of the measurements collected by OSHA inspectors while assessing compliance. The OSHA officers performed calculations on the sample results [e.g. a time-weighted average (TWA) calculated from several short-term samples] and recorded the result of their assessment in IMIS. Each record in IMIS includes information about the company in which the inspection was conducted (see Table 1). Industries are identified by a four-digit code from the 1987 or 1972 Standard Industrial Classification (SIC) and also by a six-digit code from the North American Industry Classification System (NAICS) after 1997. The description of the monitored job is entered as free text. Other characteristics of the inspection and the measurement are also recorded (Table 1). IMIS exposure data can be obtained from OSHA by any US or non-US citizen or organization through Freedom of Information Act requests for a processing fee that covers the file preparation time and with a processing time ranging from several weeks to a few months. For example, OSHA charged the investigators \$400 US to obtain IMIS data for 36 agents. OSHA also conducts health consultation aimed at helping companies improve their health and safety record. Access to this data has been limited to preserve anonymity of companies and avoid discouraging them from seeking assistance (Okun *et al*., 2004).

Literature review of previous uses of OSHA IMIS data in research

We conducted a literature review to identify scientific articles mentioning the use of measurement data collected during OSHA's enforcement activities. The search involved the keywords 'OSHA', 'IMIS', and 'occupational exposure' in PubMed. Additional references were obtained from the bibliographies of the retrieved articles. This review aimed at gathering information about the contents of IMIS, identifying the methodological approaches used to analyse these data and the challenges encountered, and collecting insights about potential biases present in this databank.

Descriptive analysis of the Chemical Exposure Health Data

Measurements from the Salt Lake Technical Center have been available from the OSHA website since May 2010 under the title 'Chemical Exposure Health Data' (http://www.osha.gov/opengov/healthsamples.html). These data can be accessed individually through search by company names, state, ZIP code, year, industry code, agent, or range of results, and downloaded as compressed XML files. The field definitions provided on the website were not complete; therefore we communicated with the Salt Lake Technical Center to define and recode all values not mentioned in the definitions. The dataset included a variable 'sampling number' that identified sequential partial-shift measurements. We used this identifier to aggregate sequential samples to calculate total sampling time, median number of samples per single evaluation as defined by a unique sampling number, and the TWA for the evaluation. When one of the samples was reported as a non-detect (i.e. concentration smaller than the limit of quantification), its value was replaced by 0 in the calculation of the average concentration. If all samples were non-detects, the aggregated value was flagged as a non-detect. The dataset also included a variable 'field number' that identified samples collected on the same sampling media. We used this to identify records belonging to a panel screen (e.g. a panel of metals), and we calculated the proportion of times an agent was quantified alone or alongside others.

Comparison of the CEHD and the IMIS data

We performed both qualitative and quantitative comparisons of IMIS and the CEHD dataset. The qualitative analysis consisted of comparing the variables and their definitions. For the quantitative analysis, we focused our comparison on lead, the most measured agent in IMIS. Analyses were restricted to the period from 1985 (1 year after the start of the CEHD dataset) to 2009 (the last year in the online CEHD data). Analysis was also restricted to personal measurements.

RESULTS

We describe our findings under each of the main study objectives: (i) a review of previous uses of OSHA data with particular focus on potential biases and statistical approaches used to interpret the data, (ii) a description of the CEHD data, and (iii) the relationship between the CEHD and IMIS lead data.

Previous uses of OSHA data in research

The literature review identified 29 publications reporting the use of OSHA measurement data, of which 26 were scientific articles; two were NIOSH reports, and a Master thesis. Most publications (18) reported the analysis of IMIS data for a single agent. Among these, silica (8) and lead (3) were the most commonly analysed. The CEHD data were not used in any scientific publications to date.

Syntheses of IMIS data were reported as early as 1983 (Oudiz *et al*., 1983) and most recently in 2011 (Hamm and Burstyn, 2011). Most publications (14) drew general portraits of exposure levels in IMIS for a pollutant or an industry/occupation. The next most frequent objective involved estimating proportions or numbers of workers exposed (4). For example, Hamm and Burstyn (2011) estimated the probability of beryllium exposure as the probability that a measurement within an industry/occupation group was higher than specified thresholds to later enable constructing a job-exposure matrix. Mendeloff (1984), Linch *et al*. (1998), and Henneberger *et al*. (2004) estimated the proportion of exposed workers within an industry using the number of workers exposed to the level recorded and total number of employees at the site. Other objectives included evaluating the potential of under-reporting measured levels in IMIS (Jones *et al*., 1986), ranking industries for exposure surveillance (Froines *et al*., 1986; Valiante *et al*., 1992), describing historical OSHA inspections (Froines, 1989), evaluating utility of IMIS in epidemiology (Stewart and Rice, 1990), assessing recording errors (Clark, 1990), identifying factors associated with exposure levels (Gómez, 1997; Melville and Lippmann, 2001), studying the effect of OSHA sampling procedures on exposure variability (Tanner-Martinez, 1997), and comparing IMIS to a French occupational exposure database (Lavoué *et al*., 2008).

Several approaches have been used to describe exposure data in IMIS: the earliest studies used descriptive univariate methods (Oudiz *et al*., 1983); the most recent ones used several multivariate statistical procedures (Table 2). These approaches can be separated into two main families: modelling a quantitative exposure level as a function of potential influential factors, or modelling the probability of an exposure level being higher than a pre-specified threshold [i.e. PEL or limit of detection (LOD)]. In the first family, linear models were generally used after logarithmic transformation of the exposure levels. In the second family, logistic or Poisson regression was used to estimate 'probability of exposure'. A common variation was to model correlation structures in the data, in particular within data measured during the same inspection (Gómez, 1997; Lavoué *et al*., 2011, 2008; Okun *et al*., 2004; Teschke *et al*., 1999). Lavoué *et al*. (2008, 2011) and Teschke *et al*. (1999) reported within-inspection correlation coefficients ranging from 0.4 to 0.7, assuming a compound symmetry in the covariance matrix.

As early as 1984, Mendeloff (1984) underlined the fact that IMIS has not been designed as an exposure surveillance tool and that the results stored within this databank could not be regarded, by default, as representative of the exposures experienced by typical workers in the USA. None of the various processes leading to the recording of an exposure level in IMIS could be considered random: industries targeted for sampling, facilities visited within an industry, occupations evaluated within a facility, workers selected within an occupation, period of time sampled, and finally recording of the measurement result into IMIS. These selection processes all potentially lead to a difference between the situations monitored by OSHA inspectors and workplace exposures experienced by the general population. Table 3 summarizes the studies that reported results related to bias in IMIS data. Most studies evaluated the relationship between exposure levels and characteristics of the company visited or of the type of inspection conducted, which may reflect differential selection of companies within an industry group (e.g. selection of 'dirtier' companies by complaint-related inspections). Because no gold standard exists, no study directly addressed whether the IMIS data represented exposure levels in the general working population (i.e. the so-called 'worst case' or 'compliance' bias). One the most

Table 1. Variables available in the IMIS and Salt Lake City OSHA laboratory electronic database.

Field	Data type	Description
Variables common to IMIS and the CEHD dataset		
Inspection number	Category	Unique identifier tied to each inspection
Establishment name	Text	Establishment name associated to inspection (names contained in the IMIS are not unique; i.e. there may be more than one variation in the way a single establishment is spelled)
City	Text	Identifies the site city where the inspection was performed
State	Category	Identifies the site state where the inspection was performed
Zip code	Category	Identifies the site zip code where the inspection was performed
SIC code	Category	Indicates the four-digit Standard Industrial Classification Code from the 1987 or 1972 version (record prior to 1987 are coded according to the 1972 system)
NAISC code	Category	North American Industrial Classification System Code (Starting in 1997)
Sampling number	Category	Unique identifier tied to single exposure assessment (there may be multiple media tied to this number in the CEHD dataset, reflecting multiple samples used for the calculation of a time-weighted sample)
Office id	Category	Unique number assigned to an OSHA Office
Date sampled	Date	Date sample was taken
IMIS substance code	Category	IMIS substance code number
IMIS Substance name	Category	Substance chemical name
Variables specific to IMIS		
State or federal	Category	Activity related to a state or federal OSHA plan
Inspection type	Category	Type of inspection: Un-programmed (complaint, referral by a safety officer, accident, follow-up, related to another inspection) Programmed (planned, related to another inspection)
Inspection coverage	Category	Comprehensive or partial survey of the establishment
Establishment size	Continuous	Number of employees in the company monitored
Employee covered	Continuous	Number of employees covered by the inspection
Employees exposed	Continuous	Number of employees in the exposure group associated with the record
Union status	Category	Union is present or not in the company monitored
Job title	Text	Short description of occupation
Frequency of exposure	Text	Short description of the frequency of exposure (e.g. 40 h per week)
Sample type	Category	Type of sample: Area, personal, blood, screening, urine, wipe, bulk
Exposure type	Category	Type of exposure: TWA, short-term exposure limit, ceiling, peak, non-detect, dose (noise), sound (noise level), not analysed, not valid
Advance notice given	Yes/no	The company was warned that an inspection would take place
Presence of employee representative	Yes/no	Employee representatives were present during the inspection
Interview of employees	Yes/no	Employees were interviewed during the inspection

Table 1. *Continued*

Field	Data type	Description
Variables specific to the CEHD data		
Instrument type	Text	Brief description of the laboratory instrument used for analysis
Lab number	Category	Unique identifier assigned by laboratory for internal use
Field number	Category	Unique identifier tied to an individual sample media submitted for analysis
Sample type	Category	Sample type of the measurement (Personal, Area, Bulk, Wipe, Screening)
Blank used	Yes/no	Sample represents an analytical blank
Time sampled	Continuous	Sample time in minutes
Air volume sampled	Continuous	Air volume sampled in liters
Sample result	Continuous	Sample result in concentration unit
Unit of measurement	Category	Unit of measurement (mg/m3, micrograms, Parts per million, milligrams, fibers/cc, percentage)
Sample weight	Continuous	Sample weight for bulks and silica samples (in mg)
Qualifier	Category	Identifies a sample as non-detect, analytical blank, approximate value, or member of a series of samples

Table 2. Multivariate techniques used to analyse IMIS exposure data.

Publication	Agent and setting	Exposure metric	Variables studied	Analytical approach
Froines *et al*. 1991	silica	median severity by inspection	industry, number of employees, union status, inspection type	logistic regression, response is inspection specific median severity being greater than 1
Gomez 1997	three subsets: lead in battery manufacturing, perchloroethylene in dry cleaning, iron oxide in welders	concentration, company specific mean concentration, probability of being greater than a specified value	job description, number of employees, union status, year, scope of inspection, type of inspection	for each dataset: linear multiple regression of company specific log-transformed mean concentrations, linear multiple regression of individual log-transformed concentrations with within inspection correlation, logistic regression of the probability for a measurement being greater than the dataset specific 75th percentile of exposure levels
Linch *et al*. 1998	silica	proportion of workers associated with a fixed severity	year, number of employees	linear model with response the transformed site specific proportion of workers exposed as a function of industry, year and number of employees
Teschke *et al*. 1999	wood dust	concentration	year, job description, number of employees, inspection type	linear multiple regression of individual log-transformed concentrations with within-company correlation

Table 2. *continued*

Publication	Agent and setting	Exposure metric	Variables studied	Analytical approach
Melville and Lippman 2001	three subsets: Asbestos abatement, toluene in auto-repair bodyshops, formaldehyde in embalmers	concentration, company specific mean concentration, probability of being greater than a specified value	job description, number of employees, union status, year, scope of inspection, type of inspection	for each dataset: linear multiple regression of company specific log-transformed mean concentrations, linear multiple regression of company specific log-transformed mean concentrations weighted by associated variances, linear multiple regression of individual log-transformed concentrations with within-inspection correlation
Coble *et al*. 2001	several agents in the pulp and paper industry	concentration	industry, job description, year	linear regression of log-transformed concentrations on year of measurement
Lurie and Wolfe 2002	hexavalent chromium	concentration, number of measurements, citations	year, industry, inspection type, inspection conducted by federal or state agency	univariate linear regressions and rank sum tests
Hennerberger *et al*. 2004	beryllium	companies with most recent inspection associated with beryllium levels greater than 0.1 or 0.5 µg/m3	year, number of employees	linear multiple regression with response the transformed site specific proportion of workers exposed as a function of SIC, year and number of employees
Middendorf 2004	noise	several noise exposure metrics	year, number of employees,	linear regression for noise exposure versus year and general linear model for noise level versus number of employees + year
Okun *et al*. 2004	lead	probability of a measurement exceeding the PEL	year, region, number of employees, union status, inspection type	SIC specific logistic regression with correlation within inspection (fit using generalized estimating equations), response is individual sample result being greater than PEL
Yassin *et al*. 2005	silica	concentration	year, industry, job description, inspection type	non parametric regression to test the hypothesis of similar mean exposure in all industries, autoregressive ARMA (2) model with errors correlated with previous and following time periods. Covariates included year, industry, and inspection type
Flanagan *et al*. 2006	silica in the construction industry	concentration	several exposure determinants not documented in IMIS+ year	linear multiple regression of individual log-transformed concentrations
Lavoué *et al*. 2008	formaldehyde	concentration	inspection type, sample type (short-term, TWA), season, industry, year, state, outside temperature,	linear multiple regression of individual log-transformed concentrations with within-inspection correlation, TOBIT models
Lavoué *et al*. 2011	formaldehyde	concentration	data source, year, sample type (short-term versus TWA), industry	linear multiple regression of individual log-transformed concentrations with within-inspection correlation, TOBIT models, multimodel inference as the model selection framework

Table 2. *continued*

Publication	Agent and setting	Exposure metric	Variables studied	Analytical approach
Hamm and Burstyn 2011	beryllium	evaluation leading to beryllium level greater that 0.1 μg/m^3 or 0.5 μg/m^3	industry, job description, measurement being TWA, year	Poisson multiple regression with random sample effect
Henn *et al*. 2011	lead	percent of samples over the PEL	industry, time period, region, number of employees, federal/state plan, union status, inspection type, advance notice of inspection, presence of employee representative, employees interviewed during inspection	logistic regression, response is the probability of a measurement being greater than the PEL

interesting observations comes from the work by Okun *et al*. (2004), who observed that the OSHA 'health consulting' data for lead had a consistently lower probability of being over the PEL compared to the 'enforcement data', albeit by a modest margin (between 1 and 5% across years) (Okun *et al*. 2004). The only comparison involving IMIS and another measurement database showed overall higher formaldehyde levels in the French database COLCHIC but similar contrasts between industries (Lavoué *et al*. 2011). Froines *et al*., (1986), and Valiente *et al*. (1992) compared how similarly industries were prioritized by IMIS, the National Occupational Exposure Survey (NOES) (Boiano and Hull, 2001), and a silicosis registry in New Jersey. They observed both similarities and discrepancies in the identified priority industries, noting that NOES was more useful as a hazard identification system, while IMIS was useful to identify overexposures for agents and industries covered by OSHA compliance activities.

Finally, some studies suggested under-reporting in IMIS. Such phenomenon implies a difference between the population of situations sampled by OSHA officers and the population of results recorded in IMIS. Jones *et al*. (1986) reviewed paper files from 451 inspections (covering 12 agents) performed in two OSHA offices between 1980 and 1983 and found that only half of the collected samples were recorded in IMIS. However, no systematic differences in median severity were found between the original inspection files and IMIS data. These figures may not be representative of the current IMIS database because the process of recording became centralized at the Salt Lake City laboratory after 1984 (Jones *et al*., 1986). In addition, the differential recording of measurements in IMIS is probably not a uniform phenomenon across OSHA offices/inspectors. For example, Mendeloff (1984) quoted an earlier study (not possible to access directly) that found that the proportion of measured exposures recorded in IMIS was higher when it corresponded with issuing a citation for overexposure.

Two particular challenges in using IMIS relate to data below the LOD. First, the status of a measurement coded as a non-detect is provided in the same variable that identifies a sample as TWA or short-term measurement ('exposure type' in Table 1). This precludes users from properly handling the non-detects because one does not know whether a non-detect was a full-shift TWA with lower LOD or a short-term sample with higher LOD. A simulation of different scenarios for the distribution of non-detects across the TWA and short-term categories for formaldehyde found non-negligible impacts on the predicted exposure levels (Lavoué *et al*., 2008). However, this characteristic would not be problematic for agents with only one type of measurement. Second, most authors reported a high percentage of non-detects in IMIS. Paraphrasing Melville and Lippman (2001), it is not possible to separate 'present but not detected' results, i.e. agent was present in the workplace but at a low level, from 'not present' results, i.e. agent was absent from the workplace. Froines *et al*. (1990) excluded not detected lead levels because zero exposure would not be a valid measure in workplaces where lead is present, thus treating them as 'not present' results. As noted by Henneberger *et al*. (2004), multiple agents are sometimes measured on the same sample media (Appendix 1). For these agents, several results may correspond to a 'not present' situation. At one extreme, if one assumes that measurements are made only when the agent was present, non-detects should be treated as censored values from an observed exposure distribution. At the other extreme, treating non-detects as 'not present' implies there is a certain prevalence of exposure across the measured industries, and that when exposure is present, the levels are those that were detectable. As a result, one would estimate the probability of exposure being present by using all data and subsequently use only detected values to estimate average exposure levels for the 'exposed' setting. Most treatments of non-detects in IMIS corresponded to one of the two interpretations: exclusion of non-detects (Freeman and Grossman, 1995; Froines *et al*., 1990; Gómez, 1997; Lurie and Wolfe, 2002; Melville and Lippmann, 2001) or inclusion of all non-detects as 'present but not detected' by replacing non-detects with values between zero and LOD (Coble *et al*., 2001; Tanner-Martinez, 1997; Teschke *et al*., 1999). Compromises are also possible. For example, Lavoué *et al*. (2011) predicted formaldehyde concentrations by including only one-third of the initial number of non-detects in their TOBIT models. Finally, some authors modelled the probability of a measurement being greater than a specified value above the LOD (Hamm and Burstyn, 2011; Henn *et al*., 2011; Henneberger *et al*., 2004; Linch *et al*., 1998; Okun *et al*., 2004).

The CEHD dataset

Prior to analysing the CEHD data, the following records were removed if they were (i) irrelevant for exposure assessment (e.g. blank samples), (ii) had uninterpretable misspellings, (iii) missing information (e.g. instrument type not provided), (iv) null values when a non-null result was expected (e.g. sampling time is 0), and (v) conflicting values (e.g. labelled a non-detect but sample results is >0). We examined all unique values of categorical variables in the dataset and assigned a standardized value when probable typing errors were identified. To facilitate the widespread use of these data, we provide a detailed description of the data cleaning process and a link to an application that recreates the cleaned data from the raw XML files on the Web in an online supplement.

The online dataset contained 1 908 373 records covering the period from 1984 to 2009; included variables are described in Table 1. To clean the data, we removed 'soil', 'gravimetric determination', and 'sample weight' measurements, which we judged not useful for exposure assessment (n=102 792). Next, we eliminated blanks (n=315 001) and records judged erroneous (n=39 705). The remaining 1 450 836 records were predominantly personal samples (78.4%), with the balance consisting of 4.3%

Table 3. Studies of IMIS exposure data having reported results related to potential biases.

Publication	Main focus	Exposure metric	Variables studied	Bias
Oudiz *et al*. 1983	silica exposures in foundries	% of exposures above PEL, severity	work area, type of foundry, number of employees	fraction of overexposures increasing with number of employees
Jones 1986	under reporting in IMIS	% of samples in OSHA reports ending up in IMIS	N.A.	slightly fewer than 50% of compliance data reported in IMIS, 25% of plants with compliance data do not appear in IMIS, under-reporting does not seem related to level of exposure
Froines *et al*. 1986**a**	general portrait of silica exposure	severity	industry, union status, inspection type, job description	despite between-industry differences, general trend of higher probability of being >PEL for complaint inspections, especially in unionized companies. No consistent trend for mean severity.
Stewart *et al*. 1990	use of IMIS for occupational epidemiologic studies	concentration	industry, job description	SIC specific measurement arithmetic mean higher for complaint inspections (median ratio of 2.4, 3 out of ten ratios less than 1)
Froines *et al*. 1990	general portrait of lead exposure	median severity by inspection	industry, number of employees, union status, inspection type	odds ratio of 3 for complaint inspections versus scheduled for the probability of a median severity within an inspection to be greater than 1

Table 3. *continued*

Publication	Main focus	Exposure metric	Variables studied	Bias
Gomez 1997	association between IMIS variables and reported exposure levels	concentration, company specific mean concentration, probability of being greater than a specified value	job description, number of employees, union status, year, scope of inspection, type of inspection	clear trend for number of employees (exposure level decrease when number of employees increase : GMs for large companies (>273 employees) are 30–40% of those from small company (<60 employees)
Tanner-Martinez 1997	effect of non-random sampling on estimation of exposure variability from IMIS data	company-specific geometric standard deviation	auto-correlation structures	GSDs smaller when estimated from few samples (n smaller than 6) or from samples within a small time period (week)
Melville and Lippman 2001	association between IMIS variables and reported exposure levels	concentration, company specific mean concentration, probability of being greater than a specified value	job description, number of employees, union status, year, scope of inspection, type of inspection	variable results. General trend of higher levels for general scope inspections. For toluene and formal-dehyde, levels associated with complaint inspections higher versus scheduled. Quantitative estimates no provided.
Lurie and Wolfe 2002	general portrait of exposure to hexavalent chromium	concentration, number of measurements, citations	year, industry, inspection type, inspection conducted by federal or state agency	greater % of non-detects in state inspections (59.8% versus 48.9%) compared to federal inspections.
Middendorf 2004	surveillance of occupational noise exposure	several noise exposure metrics	year, number of employees,	noise levels increase with number of employees (shift of 2–3 dBA from <20 to >499 employees). Mean consultation levels > mean enforcement levels (up to 4 dBA depending on year, average ~2)
Okun *et al*. 2004	trends in occupational lead exposure	probability of a measurement exceeding the PEL	year, region, number of employees, union status, inspection type	probability of being higher than PEL slightly higher for compliance data than for consultation data (between 1 and 5% across years), and for complaint inspection than for general schedule inspections (estimate of 5% from logistic regression)
Yassin *et al*. 2005	general portrait of exposure to silica dust	concentration	year, industry, job description, inspection type	programmed inspection industry specific geometric means slightly higher than overall industry specific GMs (0.077 versus 0.073 mg/m^3)
Lavoué *et al*. 2008	general portrait of expo-sure to formaldehyde	concentration	inspection type, sample type (short-term, TWA), season, industry, year, state, outside temperature,	marginal effect of inspec-tion type with complaint and referral inspections associated with slightly higher levels than scheduled inspections (7%). Exclusion of non-detects might have caused underestimation of ~20–30% for TWA results, up to 60% for short-term results.

Table 3. *continued*

Publication	Main focus	Exposure metric	Variables studied	Bias
Lavoue *et al*. 2011	comparison of formaldehyde exposure levels in IMIS and the French exposure databank COLCHIC	concentration	data source, year, sample type (short-term versus TWA), industry	formaldehyde levels somewhat higher in the French database (by 14% in average, reduced to no difference after exclusion of health sector). Contrast between most industries very similar. Exclusion of non-detects would have caused overestimation of IMIS TWA results by ~20% and underestimation of the COLCHIC short-term data by ~30%.
Henn *et al*. 2011	general portrait of exposure to lead	percent of samples over the PEL	industry, time period, region, number of employees, federal/state plan, union status, inspection type, advance notice of inspection, presence of employee representative, employees interviewed during inspection	higher probability of being over the PEL for smaller companies (1–99 versus over 500 :OR=2), federal versus state plan (OR=1.1), union versus no union (OR=1.23), advance notice of inspection (OR=1.6), absence of employee representative(OR=1.19), no employee interviewed (OR=1.33)
Teschke *et al*. 1999	exposure to wood dust for a population-based case–control study	concentration	year, job description, number of employees, inspection type	none reported in multivariate analysis. In univariate analysis, GM for planned inspection slightly lower than program related (complaint or referral, 1.86 versus 1.99 mg/m^3)

area, 7.5% wipe, 8.6% bulk, and 1.1% screening samples. Fig. 1 presents a graph of the number of measurements of all agents per year.

Of the 1082 agents, 78 agents had over 1000 personal samples. Appendix 1 presents, for these 78 agents, the sample size, percentage of non-detects, and median sample duration for the measured concentrations and time-weighted-average values, as well as the proportion of time an agent was measured as part of a panel, and the median number of agents measured on the panel when applicable.

Comparison of IMIS and the CEHD dataset

The IMIS and CEHD databases are complementary (Table 1). Specifically, IMIS provides the circumstances of measurement in the workplace but minimal sampling and analytical details. In contrast, the CEHD data provides the analytical result and associated details of the measurement. The inspection and sampling number variables were present in both datasets and a unique 'inspection number'–'sampling number' identifier was created to link the two data sets.

For lead, the extracted data from the 1985–2009 IMIS data contained 34 225 personal records, which were reduced to 33 714 records corresponding to 9905 inspections after elimination of coding errors and duplicates. The extracted CEHD lead measurements contained 73 144 analytical results, which were reduced to 48 045 time-weighted-averages corresponding to 13 916 inspections.

When the IMIS and CEHD data were merged, 23 033 (39.2%) records were in common to both IMIS and CEHD datasets ('both-IMIS-CEHD data'), 10 681 (18.2%) records were only in IMIS ('IMIS-only data'), and 25 012 (42.6%) records were only in the CEHD database ('CEHD-only data'). The distribution differed when we stratified by type of OSHA program (Table 4). Measurements collected under State OSHA plans were much less likely to be in the both-IMIS-CEHD data (13%) than measurements collected under federal OSHA (44%).

Fig. 2 shows the empirical cumulative distribution functions of the lead concentrations in each data set, with all values shown in Fig. 2a and only

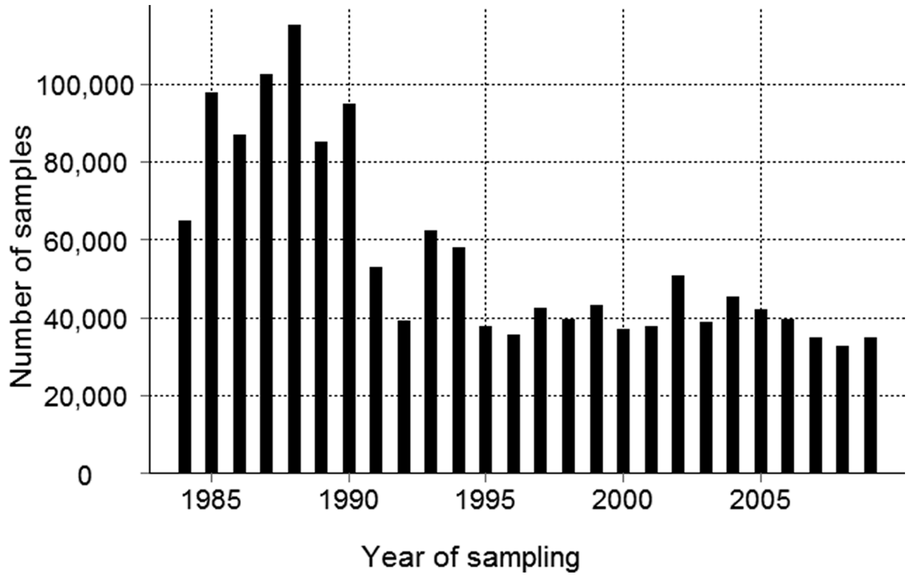

Fig. 1. Number of samples per year in the Chemical Exposure Health Data.

Table 4. Distribution of the presence of data records in IMIS-only, the Salt Lake City dataset, or both, according to the presence of OSHA stet plan.

	Federal plan		State plan [a]		Partial state plan [b]	
	n	%	*n*	%	*n*	%
IMIS only	3083	8	5948	67	1650	13
CEHD only	17 363	47	1691	19	5958	45
Both IMIS and CEHD	16 260	44	1214	14	5559	42
Total	36 706	100	8853	100	13 167	100

[a]States with OSHA state plans include: AK, AZ, CA, HI, IN, IA, KY, MD, MI, MN, NM, NC, OR, PR, SC, TN, UT, VT, VA, WA, WY.
[a]States with partial OSHA state plans include: CT, IL, NJ, NY, VI, NV.

detected values shown in Fig. 2b. As shown on the left-hand side of Fig. 2a, the percentage of non-detects differed substantially among the both-IMIS-CEHD, CEHD-only, and IMIS-only datasets (42%, 71%, and 46%, respectively), causing very different empirical cumulative distribution functions. On the other hand, Fig. 2b shows very similar empirical cumulative distribution functions when data is restricted to detected samples. Hence, the 25th percentile, median, and 75th percentile were similar among the both-IMIS-CEHD, CEHD-only and IMIS-only datasets for detected samples (25th percentile: 0.015, 0.015, 0.010; median: 0.042, 0.043, 0.039; 75th percentile: 0.150, 0.141, 0.150). Values for the both-CEHD-IMIS subset in the previous calculation data were taken from the IMIS results data; there were negligible differences when the CEHD values were used.

To illustrate the potential implications of the 'not present' versus 'not detected' issue mentioned above, we considered the IMIS-only dataset for lead (46% of non-detects). If the non-detects were primarily collected in locations where lead exposure was not present, excluding the non-detects would yield a geometric mean (GM) of 0.042 mg/m^3 (geometric standard deviation, GSD=9.2). If the non-detects were 'present but not detected', including the non-detects using an imputation based on regression on order statistics (Helsel, 2005) would yield a GM of 0.007 mg/m^3 (GSD=18.5), based on a LOD of 0.00284 mg/m^3 (OSHA method ID125 and the median sampling time in the CEHD lead data, 222 min.)

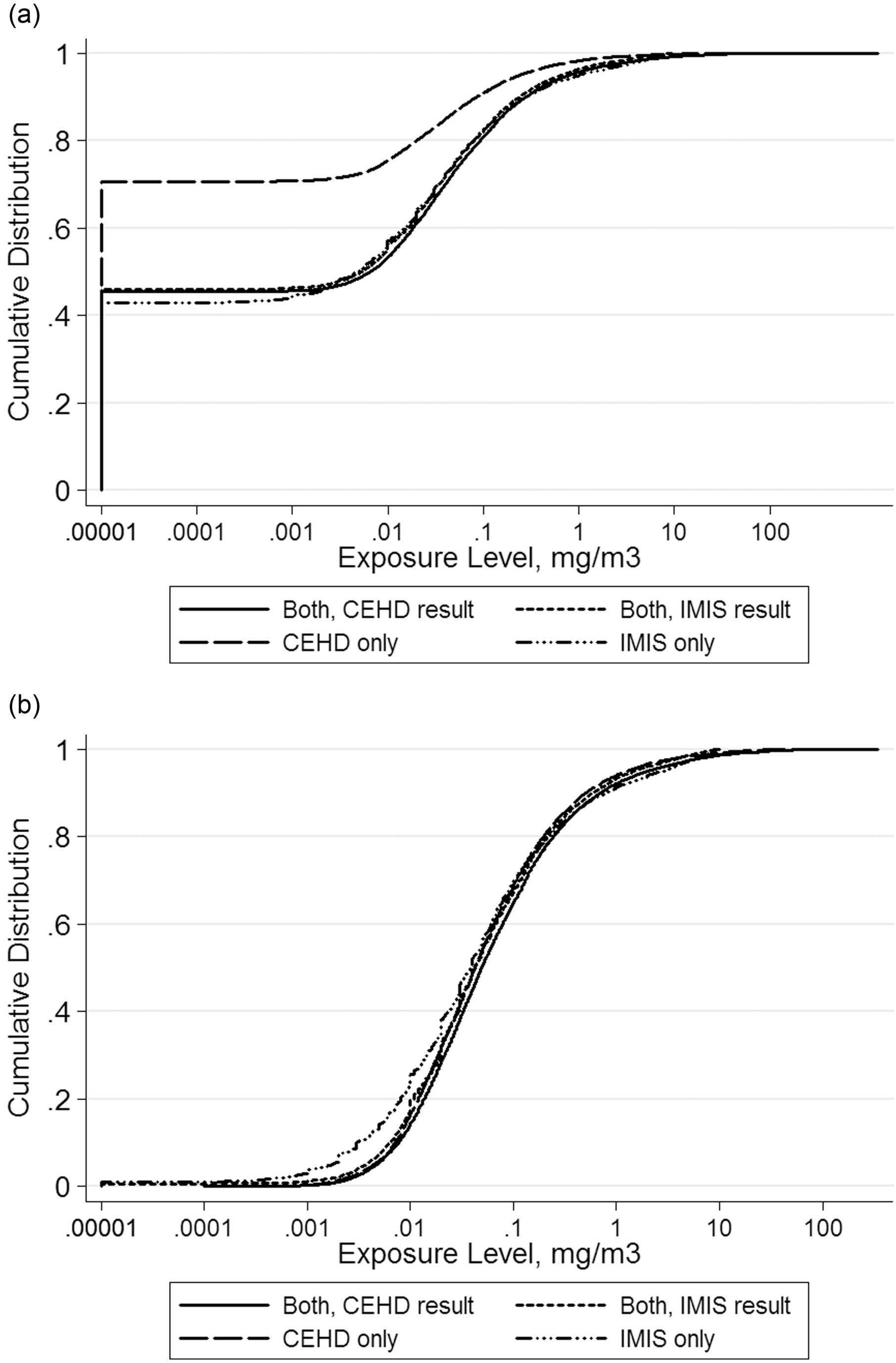

Fig. 2. (a) Empirical cumulative distribution functions of the lead concentrations in the IMIS-only, CEHD-only, and both-IMIS-CEHD datasets. (b) Empirical cumulative distribution functions of the detected lead concentrations in the IMIS-only, CEHD-only, and both-IMIS-CEHD datasets.

DISCUSSION

Stewart and Rice (1990) were among the first to describe the potential of IMIS as a source of exposure information for exposure assessment in epidemiology. Their recommendations reflected the small number of records in IMIS at that time. After more than two decades of sampling activities by OSHA, the now over 1 million personal measurements recorded, and the recent public release of a complementary CEHD data, the present work provides a timely update to Stewart and Rice's initial portrait.

The descriptive analysis of the CEHD data showed that the majority of results corresponded to measurements in the breathing zone of workers. Close to 80 agents were associated with >1000 personal samples over the period of 1984 to 2009. This dataset currently represents one of the largest public sources of retrospective multi-industry exposure information. The freely available software accompanying this manuscript, which automatically recreates the dataset summarized in Appendix 1, should facilitate its widespread access by the researchers.

The comparison of variables in the CEHD and IMIS databanks shows that it is important to link both datasets to take full advantage of the available ancillary information. The CEHD data supplements the IMIS data with the sampling duration, analytical method, and presence of other substances on the same sampling media. However, based on the example of lead, only 40% of the data is included in both datasets. The IMIS-only data may be explained by measurements analysed at other laboratories. The CEHD-only data may reflect an under-reporting of samples into IMIS, supporting previous comments by Mendeloff (1984) and Jones *et al*. (1986). Moreover, the proportion of non-detects in the CEHD lead data was significantly higher than in the both-IMIS-CEHD and IMIS-only datasets, supporting the hypothesis that the IMIS under-reporting is differential: non-detects seem less likely to be recorded in IMIS than other samples. Detected values, on the other hand, had similar empirical cumulative distribution functions in the IMIS-only, CEHD-only, and both-IMIS-CEHD datasets, suggesting that differential reporting only affects non-detects. Taken alone, the value of the CEHD data for exposure assessment may appear less than that of IMIS, because of the very limited ancillary information (only industry is provided). However, the CHED dataset offers a unique opportunity to explore biases in the OSHA measurement data. The comparisons between the two databanks presented here provides preliminary insights into the strengths and limitations of both data sets, but more comprehensive analyses are required.

Most commentators agree that the IMIS data cannot be regarded by default as providing representative portrait of workplace exposure in the USA. While it is straightforward to use IMIS and the CEHD data to identify instances of overexposure, estimating average exposure conditions from these sources is challenging given the number of potential biases (i.e. selection of industries, companies, workers, high, or low exposure situations). However, many authors reported temporal trends in exposures estimated from IMIS data that were compatible with other sources of data, implying that at least extrapolation of relative time trends from these data may be reliable. The critical issue, given the paucity of exposure data in general, is whether these data are useful despite the potential for bias. Bias in IMIS has mostly been studied internally by evaluating association between reported levels and information on the circumstances associated with an inspection, such as the reason for the inspection, interview of employees, or on the company itself, such as company size or the presence of a union. To date, no bias of appreciable magnitude has been consistently reported across studies and agents. Moreover, biases linked to these variables can be adjusted for in multivariate models. Regarding the differential selection of occupations within a company, the IMIS variable 'job description', if it was standardized, would assist in addressing this bias since one would know to what occupations the measured levels are relevant. Some authors have manually recoded this variable when their dataset was restricted to few industries (Teschke *et al*., 1999; Hamm and Burstyn, 2011). More recently, Slutsky *et al*. developed an algorithm to automatically create standard occupations across all industries in IMIS from the text description (Slutsky *et al*., 2011). The analysis of variables internal to IMIS, while informative, cannot evaluate adequately the relationship between exposure levels in IMIS and those occurring in US workplaces. Although tests of external validity by Okun *et al*. (2004), and Lavoué *et al*. (2011) are encouraging, more external validation efforts are needed. No study has directly addressed the issue of differential under-reporting of non-detects in IMIS. This phenomenon could affect both analyses of average exposure levels and the probability of a measurement being higher than some threshold and might well have hampered the discovery of biases related to the variables mentioned above. The possibility of using IMIS data in non-US settings has only be assessed in only one study (Lavoué *et al*. 2011). Despite this encouraging insight, transportability of IMIS should not be assumed by default without further comparison exercises. Finally, the inability to identify repeated measurements on workers in IMIS precludes its use for formal assessment of individual overexposure as defined by Tornero-Velez *et al*. (1997).

The interpretation of non-detects in OSHA data as 'not present' or 'present but not detected' is important given the high percentages of recorded non-detects in both the CEHD and IMIS data. These high values data suggest that reality may well lie closer to the 'not present' interpretation. However, little empirical evidence is available, and this phenomenon may well prove to be context specific rather than general. Recent advances in mixture modelling, by allowing the simultaneous estimation of prevalence of exposure and average levels when exposure is present, represent a promising avenue to address this issue, which is of particular interest for studies aiming at estimating the numbers of workers exposed above certain level (Chu *et al*., 2008; Taylor *et al*., 2001).

The potential under-reporting of non-detects measured by OSHA inspectors further complicates the interpretation of IMIS exposure results. We believe important advances can be made if we better understand the mechanisms by which (i) records with non-detectable values arise (e.g. studying cases where multiple agents are assessed on the same media) and (i) data is reported to IMIS (e.g. by identifying determinants of under-reporting based on CEHD/IMIS comparisons across agents, industries, and periods).

In conclusion, the combination of the IMIS data and the CHED data from the Salt Lake City OSHA laboratory probably forms the largest source of multi-industry exposure data in North America. While they contain complementary information, the two datasets only partially overlap. The lack of empirical information about biases and the interpretation and treatment of non-detects constitute the biggest challenges to the use of OSHA measurement data for assessing exposure in the general population, in particular because of potential differential under-reporting of measurements into the IMIS databank. Hence, while IMIS can in principle be used for identifying high exposure situations and assessing relative time trends, further work is needed to evaluate more comprehensively its use for estimating average exposure levels and estimating numbers of workers exposed, as well as assessing its transportability to international settings. We believe these hurdles should not deter researchers from using the IMIS/CEHD data, especially since most sources of exposure information are plagued with similar problems. Based on our own experience with the OSHA measurement data and the presented literature review, we offer the following recommendations to future users:

- Use both the CEHD and IMIS data because they complement each other.
- Use multivariable analysis tools to account for possible associations with ancillary information.
- Account for correlation within-inspection and within-company because it may affect estimates of variability and main effects.
- Create a standardized occupation code, as occupations have often shown to be better predictors of exposure than industry, and share with scientific community at large dictionaries/algorithms that translate free-text job descriptions to standard codes.
- The generally high proportion of non-detects indicates that simple imputation methods should be avoided and methodological research to address this challenge should be encouraged as none of the methods used in the past are entirely satisfactory.
- Perform sensitivity analyses to assess the potential impact of differential under-reporting in IMIS, including separate analyses of the IMIS-only, CEHD-only, and common datasets.

Acknowledgements—We are thankful to Nardin Rezk and Jennifer Yu for background work on the CEHD data, to Warren Hendrick from the Salt Lake Technical Center for answering our questions about the CEHD data, to Sarah Locke for background work on the IMIS data, and to the OSHA Directorate of Information Technology for providing the IMIS data. JL was supported by the Canadian Cancer Society Research Institutes, the Fond de la Recherche en Santé du Québec, and the Réseau de Recherche en Santé Environnementale. MF was supported by the Intramural Research Program of the National Cancer Institute's Division of Cancer Epidemiology and Genetics.

APPENDIX 1

Descriptive statistics of the Salt Lake City personal samples, limited to agents with >1000 samples

Chemical family	Name	Sample size	Not measured alone (%) (A)	Other agents (B)	Non-detects (%)	Median duration (min)	Number of evaluations (C)	Median sample number per evaluation (D)	Median duration per evaluation (min) (E)
solvent	Acetone	9508	79	2	24	60	2674	3	286
solvent	Benzene	5216	73	3	76	68	1732	2	282
solvent	2-Butanone	16 496	27	2	49	60	3445	4	315
solvent	2-Butoxyethanol	3001	36	2	40	90	1253	2	297
solvent	n-Butyl Acetate	7117	94	3	29	60	2236	2	248
solvent	n-Butyl Alcohol	4039	93	3	45	73	1385	2	290
solvent	Diacetone Alcohol	1130	83	3	50	60	307	3	307
solvent	2-Ethoxyethyl Acetate	1157	83	3	53	86	427	2	278
solvent	Ethyl Acetate	3734	93	3	28	60	1034	3	270
solvent	Ethyl Alcohol	3236	75	3	38	52	855	3	239
solvent	Ethyl Benzene	5042	99	3	31	72	1812	2	264
solvent	Heptane (n-Heptane)	1674	80	3	28	60	446	3	310
solvent	Hexane (n-Hexane)	3323	85	2	23	59	920	3	249
solvent	Hexone	6901	95	3	34	70	2117	3	304
solvent	Isobutyl Acetate	1763	98	3	32	69	522	3	344
solvent	Isobutyl Alcohol	1525	97	4	40	60	421	3	310
solvent	Isopropyl Alcohol	8476	77	3	24	60	2429	3	267
solvent	Methyl Alcohol	1887	33	1	46	60	496	3	274
solvent	Methyl (n-amyl) ketone	1665	95	3	42	63	554	3	247
solvent	Methyl Chloroform	5321	45	2	20	56	1569	2	216
solvent	Methylene Chloride	8717	30	2	32	52	2630	2	205
solvent	Tetrachloroethylene (Perchloroethylene)	4462	29	2	18	50	1430	2	201
solvent	Petroleum Distillates (Naphtha) (Rubber Solvent)	6812	84	2	53	64	2172	2	266
solvent	Phenol	1397	4	2	49	181	737	2	395
solvent	n-Propyl Alcohol	1073	86	3	28	56	246	4	346
solvent	n-Propyl Acetate	1273	96	2	16	60	314	4	362
solvent	Stoddard Solvent	9514	71	2	44	60	2837	3	243

Appendix 1. *continued*

Chemical family	Name	Sample size	Not measured alone (%) (A)	Other agents (B)	Non-detects (%)	Median duration (min)	Number of evaluations (C)	Median sample number per evaluation (D)	Median duration per evaluation (min) (E)
solvent	Toluene	29 767	88	2	15	60	9331	2	255
solvent	Trichloroethylene	2577	35	1	17	59	819	2	217
solvent	Trimethylbenzene (mixed isomers)	1542	90	3	33	75	565	2	224
solvent	Xylene	24 392	93	2	22	66	8274	2	255
solvent	VM and P Naphtha	4251	80	3	50	52	1215	2	197
metal	Antimony and Compounds (as Sb)	50 186	100	12	96	224	34 345	1	410
metal	Beryllium and Beryllium Compounds (as Be)	49 639	100	12	96	224	33 981	1	410
metal	Cadmium Dust (as Cd)	1105	97	3	42	190	670	1	399
metal	Cadmium Fume (as Cd)	24 558	100	13	89	215	16 305	1	413
metal	Chromium, Metal, and Insoluble Salts	50 206	100	12	63	221	34 268	1	409
metal	Chromic Acid and Chromates (as CrO3)	4518	64	1	56	154	2970	1	345
metal	Chromium (VI) – TWA	1789	84	1	49	240	1310	1	410
metal	Cobalt, Metal, Dust, and Fume (as Co)	49 089	100	12	91	224	33 590	1	410
metal	Copper Dusts and Mists (as Cu)	1254	93	3	20	238	879	1	420
metal	Copper Fume (as Cu)	50 995	100	12	37	224	34 938	1	410
metal	Iron Oxide Fume	49 303	100	12	15	223	33 681	1	410
metal	Lead, Inorganic (as Pb)	73 144	89	12	60	222	50 362	1	408
metal	Manganese Fume (as Mn)	49 020	100	12	36	223	33 452	1	410
metal	Mercury (Vapor) (as Hg)	1310	47	1	21	160	681	1	412
metal	Molybdenum (as Mo), Insoluble Compounds (Total Dust)	48 386	100	12	91	224	33 064	1	410
metal	Nickel, Metal, and Insoluble compounds (as Ni)	49 427	100	12	77	224	33 834	1	410

Appendix 1. *continued*

Chemical family	Name	Sample size	Not measured alone (%) (A)	Other agents (B)	Non-detects (%)	Median duration (min)	Number of evaluations (C)	Median sample number per evaluation (D)	Median duration per evaluation (min) (E)
metal	Silver, Metal, and Soluble Compounds (as Ag)	2160	94	6	48	213	1483	1	389
metal	Tin, inorganic compounds (except oxides) (as Sn)	2216	85	6	78	240	1689	1	400
metal	Vanadium fume (as V2O5)	48 748	100	12	92	224	33 302	1	410
metal	Zinc Oxide Fume	50 346	100	12	37	223	34 439	1	410
metal	Cadmium	20 428	99	12	83	227	14 176	1	406
gas	Ammonia	1808	22	1	35	119	911	2	300
gas	Ethylene Oxide	1901	2	2	53	35	738	2	240
Gas	Formaldehyde	9063	2	1	38	111	4699	2	240
Gas	Hydrogen Chloride	2205	37	1	66	15	925	2	45
Gas	Sulfur Dioxide	1572	29	1	35	117	597	2	410
Gas	Vinyl Chloride	2126	4	1	86	45	401	4	285
isocyanates	Methylene bisphenyl isocyanate	7009	35	3	75	15	3325	2	30
isocyanates	Hexamethylene Diisocyanate	3660	56	3	70	18	1631	2	45
isocyanates	Toluene-2,4-Diisocyanate (TDI)	4455	93	2	77	17	1981	2	45
isocyanates	Toluene-2,6-Diisocyanate	3969	99	2	69	17	1785	2	45
PAHs	Chrysene	1623	100	2	69	240	1261	1	390
PAHs	Coal Tar Pitch Volatiles (benzene soluble fraction)	1707	72	2	28	230	1341	1	376
PAHs	Naphtha (Coal Tar)	1299	87	2	58	70	408	3	281
PAHs	Benzo [a] Pyrene	1632	99	2	78	240	1274	1	392
other dust / fibers	Fluorides (as F)	991	35	1	62	103	490	1	339
other dust / fibers	Silica, Crystalline Quartz (Respirable Fraction)	25 230	32	1	50	242	19 433	1	400

Appendix 1. *continued*

Chemical family	Name	Sample size	Not measured alone (%) (A)	Other agents (B)	Non-detects (%)	Median duration (min)	Number of evaluations (C)	Median sample number per evaluation (D)	Median duration per evaluation (min) (E)
other dust / fibers	Silica, Crystalline Cristobalite, Respirable Dust	3248	88	1	98	267	2419	1	421
other dust / fibers	Asbestos (all forms)	16 847	1	2	59	93	9015	1	218
other dust / fibers	Particulates not otherwise regulated (Respirable Fraction)	35 123	91	13	84	220	24 142	1	409
other dust / fibers	Particulates not otherwise regulated (Total Dust)	18 513	49	1	35	213	12 744	1	403
other dust / fibers	Silica (Quartz, Total)	140	68	1	92	180	109	1	270
other	Arsenic, Inorganic	7209	94	2	51	223	4812	1	415
other	Cyclohexanone	1337	24	2	30	83	444	2	330
other	Methyl Methacrylate	1269	17	1	47	61	358	3	310
other	Nitric Acid	938	69	1	71	141	603	1	336
other	Sodium Hydroxide	1446	28	1	38	135	984	1	260
other	Styrene	13 732	25	1	11	60	3943	3	320
other	Sulfuric Acid	1500	54	1	54	206	1084	1	348

[1] http://www.osha.gov/opengov/healthsamples.htm

REFERENCES

Boiano JM, Hull RD. (2001) Development of a National Occupational Exposure Survey and Database associated with NIOSH hazard surveillance initiatives. Appl Occup Environ Hyg; 16: 128–34.

Burns DK, Beaumont PL. (1989) The HSE National Exposure Database–(NEDB). Ann Occup Hyg; 33: 1–14.

Chu H, Nie L, Kensler TW. (2008) A Bayesian approach estimating treatment effects on biomarkers containing zeros with detection limits. Stat Med; 27: 2497–508.

Clark NJ. (1990) Validation of sampling data from the Occupational Safety and Health Administration (OSHA) Integrated Management Information System (IMIS). Am Ind Hyg Assoc J; 51: A799.

Coble JB, Lees PS, Matanoski G. (2001) Time trends in exposure measurements from OSHA compliance inspections of the pulp and paper industry. Appl Occup Environ Hyg; 16: 263–70.

Creely KS, Cowie H, Van Tongeren M *et al.* (2007) Trends in inhalation exposure–a review of the data in the published scientific literature. Ann Occup Hyg; 51: 665–78.

Freeman CS, Grossman EA. (1995) Silica exposures in workplaces in the United States between 1980 and 1992. Scand J Work Environ Health; 21 Suppl 2: 47–9.

Froines JR. (1989) Worksite inspection and the control of occupational disease. The OSHA experience. Ann N Y Acad Sci; 572: 177–83; discussion 221–3.

Froines JR, Baron S, Wegman DH *et al.* (1990) Characterization of the airborne concentrations of lead in U.S. industry. Am J Ind Med; 18: 1–17.

Froines JR, Dellenbaugh CA, Wegman DH. (1986) Occupational health surveillance: a means to identify work-related risks. Am J Public Health; 76: 1089–96.

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.

Gomez MR. (1993). A proposal to develop a national occupational exposure databank. App. Occup. Environ. Hyg.; 8(9): 768–774.

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

Hamm MP, Burstyn I. (2011). Estimating occupational beryllium exposure from compliance monitoring data. Archives of Environmental & Occupational Health; 66(2): 75–86.

Hein MJ, Waters MA, Ruder AM *et al.* (2010) Statistical modeling of occupational chlorinated solvent exposures for case-control studies using a literature-based database. Ann Occup Hyg; 54: 459–72.

Hein MJ, Waters MA, van Wijngaarden E *et al.* (2008) Issues when modeling benzene, toluene, and xylene exposures using a literature database. J Occup Environ Hyg; 5: 36–47.

Helsel D. (2005). Non detects and data analysis - Statistics for censored environmental data. Hoboken, NJ: John Wiley & Sons, Inc.

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.

Henneberger PK, Goe SK, Miller WE *et al.* (2004) Industries in the United States with airborne beryllium exposure and estimates of the number of current workers potentially exposed. J Occup Environ Hyg; 1: 648–59.

Koppisch D, Schinkel J, Gabriel S *et al.* (2012) Use of the MEGA exposure database for the validation of the Stoffenmanager model. Ann Occup Hyg; 56: 426–39.

Kromhout H, Vermeulen R. (2000) Long-term trends in occupational exposure: Are they real? What causes them? What shall we do with them? Ann Occup Hyg; 44: 325–7.

Jones CA, Weld L, Gray W, Greenlee P, Quinn M, Wiarda E. (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 (NTIS No. PB2003-104588).

Kauffer E, Vincent R. (2007) Occupational exposure to mineral fibres: analysis of results stored on colchic database. Ann Occup Hyg; 51: 131–42.

Kauppinen T. (2001) Finnish occupational exposure databases. Appl Occup Environ Hyg; 16: 154–8.

Lavoué J, Gérin M, Vincent R. (2011) Comparison of formaldehyde exposure levels in two multi-industry occupational exposure databanks using multimodel inference. J Occup Environ Hyg; 8: 38–48.

Lavoue J, Vincent R, Gerin M. (2008) Formaldehyde exposure in U.S. industries from OSHA air sampling data. J Occup Environ Hyg; 5: 575–87.

Lenvik K, Osvoll PO, Woldbaek T. (1999) Occupational exposure to styrene in Norway, 1972–1996. Appl Occup Environ Hyg; 14: 165–70.

Linch KD, Miller WE, Althouse RB *et al.* (1998) Surveillance of respirable crystalline silica dust using OSHA compliance data (1979–1995). Am J Ind Med; 34: 547–58.

Liu S, Hammond SK, Rappaport SM. (2011) Statistical modeling to determine sources of variability in exposures to welding fumes. Ann Occup Hyg; 55: 305–18.

Lurie P, Wolfe SM. (2002). Continuing exposure to hexavalent chromium, a known lung carcinogen: An analysis of OSHA compliance inspections, 1990–2000. Am. J. Ind. Med.; 42(5): 378–383. John Wiley & Sons.

Melville R, Lippmann M. (2001). Influence of Data Elements in OSHA Air Sampling Database on Occupational Exposure Levels. App. Occ. Env. Hyg.; 16(9): 884–899.

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

Middendorf PJ. (2004) Surveillance of occupational noise exposures using OSHA's Integrated Management Information System. Am J Ind Med; 46: 492–504.

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.

Oudiz J, Brown JW, Ayer HE *et al.* (1983) A report on silica exposure levels in United States foundries. Am Ind Hyg Assoc J; 44: 374–6.

Park D, Stewart PA, Coble JB. (2009) Determinants of exposure to metalworking fluid aerosols: a literature review and analysis of reported measurements. Ann Occup Hyg; 53: 271–88.

R Development Core Team. (2008). R: A Language and Environment for Statistical Computing. Vienna Austria R Foundation for Statistical Computing. R Foundation for Statistical Computing. Retrieved from http://www.r-project.org

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.

Slutsky A, An Y, Hu T, Burstyn I. (2011). Automatic Approaches to Clustering Occupational Description Data for. Prediction of Probability of Workplace Exposure to Beryllium. The 2011 IEEE International Conference on Granular Computing. Taiwan, November 21–23. pp 596–601; ISBN: 978-1-4577-0372-0 (doi:10.1109/GRC.2011.6122664).

Stewart PA, Rice C. (1990). A source of exposure data for occupational epidemiology studies. App. Occ. Env. Hyg.; 5(6): 359–363.

Symanski E, Kupper LL, Rappaport SM. (1998) Comprehensive evaluation of long-term trends in occupational exposure: Part 1. Description of the database. Occup Environ Med; 55: 300–9.

Tang TK, Siang LH, Koh D. (2006) The development and regulation of occupational exposure limits in Singapore. Regul Toxicol Pharmacol; 46: 136–41.

Tanner-Martinez, L. (1997). Effect of non random sampling on estimation of exposure variability from IMIS data. University of Alabama.

Taylor DJ, Kupper LL, Rappaport SM *et al.* (2001) A mixture model for occupational exposure mean testing with a limit of detection. Biometrics; 57: 681–8.

Teschke K, Marion SA, Vaughan TL *et al.* (1999) Exposures to wood dust in U.S. industries and occupations, 1979 to 1997. Am J Ind Med; 35: 581–9.

Tornero-Velez R, Symanski E, Kromhout H, Yu RC, Rappaport SM. (1997). Compliance versus risk in assessing occupational exposures [published erratum appears in Risk anal 1997 Oct;17(5):657]. Risk Analysis; 17(3): 279–292.

U.S. Congress. (1970). Public Law 91–596: Occupational Safety and Health Act of 1970. 84 STAT. 1590, 91st Congress, S.2193. Senate and House of Representatives of the United States of America, Washington, DC.

Valiante DJ, Richards TB, Kinsley KB. (1992) Silicosis surveillance in New Jersey: targeting workplaces using occupational disease and exposure surveillance data. Am J Ind Med; 21: 517–26.

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

Vinzents PS, Carton B, Fjeldstad P, Rajan B, Stamm R. (1995). Comparison of exposure measurements stored in European databases on occupational air pollutants and definitions of core information. App. Occ. Env. Hyg.; 10(4): 351–354.

Yassin A, Yebesi F, Tingle R. (2005) Occupational exposure to crystalline silica dust in the United States, 1988–2003. Environ. Health Perspect.; 113: 255–60.