Université de Montréal

Caractérisation des mesures d'exposition recueillies par l'agence fédérale américaine OSHA pour l'estimation des expositions professionnelles en Amérique du Nord

par

Philippe Sarazin

Département de Santé Environnementale et Santé au Travail

École de Santé Publique

Thèse présentée à l'École de Santé Publique

en vue de l'obtention du grade de docteur

en Santé Publique

de l'option Toxicologie et Analyse du Risque

Juin, 2016

Université de Montréal

Faculté des études supérieures et postdoctorales

Cette thèse intitulée :

Caractérisation des mesures d'exposition recueillies par l'agence fédérale américaine OSHA pour l'estimation des expositions professionnelles en Amérique du Nord

Présentée par :

Philippe Sarazin

a été évaluée par un jury composé des personnes suivantes :

Dre Audrey Smargiassi, présidente-rapporteur

Dr Jérôme Lavoué, directeur de recherche

Dre Geetanjali Datta, membre du jury

Dr Renaud Persoons, examinateur externe

Dr Sébastien Sauvé, représentant du doyen

Résumé

La banque de données IMIS (Integrated Management Information System) de l'agence américaine OSHA (Occupational Safety and Health Administration) contient l'ensemble des mesures de l'exposition effectuées par les inspecteurs d'OSHA chargés de vérifier la conformité aux valeurs limites d'exposition. Les résultats analytiques correspondant aux prélèvements effectués par les inspecteurs sont également disponibles dans la banque CEHD (Chemical Exposure Health Data). Ces deux banques représentent une source d'information potentielle majeure sur les conditions d'exposition aux substances chimiques en Amérique du Nord. Cependant, leur représentativité par rapport à la distribution réelle des niveaux d'exposition retrouvés dans les milieux de travail est largement inconnue. L'objectif de cette thèse est d'établir dans quelle mesure les données de contamination de l'air recueillies par l'agence fédérale américaine OSHA peuvent être utilisées pour l'estimation des expositions professionnelles en Amérique du Nord.

Les analyses ont porté sur 511 047 et 588 818 mesures d'exposition contenues dans les banques IMIS et CEHD respectivement, pour la période 1979-2011. Premièrement, des modèles additifs généralisés ont été utilisés pour étudier l'association entre les variables reflétant les caractéristiques des établissements visités et des inspections et les niveaux d'exposition pour 77 agents chimiques (90% du contenu d'IMIS). Dans un second temps, une approche de régression de Poisson modifiée a été utilisée pour étudier les facteurs déterminants l'enregistrement ou non des échantillons de CEHD dans la banque IMIS en jumelant les deux banques pour 78 agents chimiques. Finalement, des modèles CART (Classification And Regression Tree) ont été développés permettant de prédire, parmi les résultats non détectés de la banque IMIS, lesquels correspondent à des mesures courte durée

ou des moyennes pondérées sur 8 heures (VEMP-8h) en se basant sur les variables communes aux banques IMIS et CEHD.

Dans la première analyse, les modèles statistiques ont montré que les niveaux d'exposition étaient plus susceptibles de dépasser la TLV (threshold limit value) pour les mesures effectuées sous un régime OSHA fédéral par rapport au régime OSHA d'État (rapport de cote (RC) de 1,22 à travers les agents). La probabilité de dépasser la TLV augmentait avec le nombre total des amendes reçues par un établissement, indépendamment de la nature des infractions (RC de 1,54 à travers les agents entre les catégories « élevée » et « aucune »). Elle était également plus élevée pour les visites de suivi que pour les visites planifiées (RC de 1,61). Dans la deuxième analyse, la comparaison des banques IMIS et CEHD a montré un taux d'enregistrement global de 38% des données CEHD dans IMIS. Les résultats non détectés (particulièrement ceux mesurés sur un panel d'agents – p. ex. panel de métaux) étaient moins susceptibles d'être enregistrés dans IMIS (risque relatif ~0,6). Finalement, les modèles CART ont prédit plus précisément le type de prélèvement (courte durée, VEMP-8h) pour les résultats non détectés dans IMIS que des méthodes simples d'attribution (p. ex. attribution du type le plus fréquent parmi les résultats détectés) pour les agents les plus pertinents (c.-à-d. ceux ayant une proportion substantielle de mesures ND, courte durée et VEMP-8h).

Nos résultats ont montré la présence de plusieurs mécanismes de sélection dans le processus conduisant à l'enregistrement d'une mesure d'exposition dans IMIS, ce qui suggère l'existence de différences systématiques entre les niveaux rapportés dans les banques OSHA et les niveaux moyens d'exposition dans la population de travailleurs. La prise en compte des informations contextuelles aux mesures et l'emploi de méthodes prédictives peuvent aider à

pallier partiellement ces biais et ainsi raffiner les portraits d'exposition établis à partir des données d'OSHA.

Mots-clés : IMIS; CEHD; banques de données d'exposition professionnelle; modèles statistiques d'exposition; méta-analyse; CART; stratégie de mesure; évaluation de l'exposition; hygiène du travail

Abstract

The Integrated Management Information System (IMIS) contains exposure measurements taken by the U.S. Occupational Safety and Health Administration (OSHA) inspectors to verify compliance with permissible exposure limits. Supplementary data containing analytical results of the field samples are available in the Chemical Exposure Health Database (CEHD). These databanks represent a major potential source of information on exposure conditions in North American workplaces. However, the degree to which they represent the actual distribution of the exposure levels found in the workplace is largely unknown.The objective of this thesis is to examine the extent to which exposure data collected by OSHA can be used for estimating occupational exposure in North America.

Analyses focused on 511 047 and 588 818 exposure measurements in IMIS and CEHD respectively, for the period 1979-2011. First, generalized additive models were used to explore associations between exposure levels in IMIS and ancillary variables reflecting characteristics of establishments and inspections for 77 chemical agents (90% of IMIS content). Second, modified Poisson regression was used to identify determinants of recording or not of CEHD samples in IMIS by linking both databanks for 78 agents. Finally, Classification And Regression Tree (CART) models were applied to predict which non-detected (ND) results stored in IMIS are 8-hour time-weighted average (TWA) or short-term samples, based on common variables available in IMIS and CEHD databanks.

In the first analysis, statistical modelling showed that measurements collected under federal OSHA plans were more likely to have a sample result exceed the TLV compared to measurements collected under state OSHA plans (odds ratio (OR) of 1,22 across agents). An

increase in the total amount of penalty assessed to a company was associated with higher odds of having a sample result exceed the TLV (OR of 1,54 across agents for « high » vs. « none »). Follow-up inspections were more likely to have a sample result exceed the TLV compared to planned inspections (OR of 1,61 across agents). In the second analysis, linkage between CEHD and IMIS showed a 38% overall proportion of CEHD samples recorded into IMIS. Non-detects (especially ND records corresponding to analytical panels – e.g. panel of metals) were less likely to be recorded in IMIS (relative risk ~0,6). Finally, CART models predicted more accurately which IMIS ND results were TWA or short-term samples compared to simple methods of assignment (e.g. assignment of the most frequent category from detected values) for the most relevant agents (i.e. with high proportions of ND, short-term, and TWA results).

Our findings showed the presence of several selection mechanisms in the process leading up to the recording of a sample in IMIS, which suggest systematic differences exist between OSHA measurements and actual occupational exposures in the general U.S. working population. These biases can be partially controlled by using ancillary information on exposure measurements together with predictive methods, thus helping to draw more accurate portraits of exposure levels from OSHA data.

Keywords: IMIS; CEHD; occupational exposure databanks; statistical models of exposure; meta-analysis; CART; measurement strategy; exposure assessment; industrial hygiene

Table des matières

Liste des tableaux

Chapitre 1

Chapitre 2

Chapitre 3

Chapitre 4

Chapitre 5

Chapitre 6

Liste des figures

Chapitre 1

Chapitre 2

Chapitre 3

Chapitre 4

Chapitre 5

Chapitre 6

Liste des sigles et des abréviations

En français

BDEP: Banques de données d'exposition professionnelle

CIRC: Centre International de Recherche sur le Cancer

GAM: Modèles additifs généralisés

GLM: Modèles linéaires généralisés

HAP: Hydrocarbures aromatiques polycycliques

INRS: Institut National de Recherche et de Sécurité

IRSST: Institut de recherche Robert-Sauvé en santé et en sécurité du travail

RC: Rapport de cotes

RIE: Indices relatifs d'exposition

RR: Risque relatif

SST: Santé et sécurité au travail

VECD: Valeur d'exposition de courte durée

VEMP: Valeur d'exposition moyenne pondérée

VLE: Valeur limite d'exposition

En anglais

AIC: Akaike information criterion

ART: Advanced REACH Tool

CART: Classification And Regression Tree

CEHD: Chemical Exposure Health Data

CI: Confidence interval

CP: Complexity parameters

CT: Classification tree

DOL: Department of Labor

FOIA: Freedom of Information Act

GAM: Generalized additive model

GAMM: Generalized additive mixed model

IDLH: Immediately dangerous for life and health

IMIS: Integrated Management Information System

IQR: Interquartile range

LOD: Limit of detection

NAICS: North American Industry Classification System

ND: Non-detect

NEDB: National Exposure Database

OEDB: Occupational exposure databank

OIS: OSHA Information System

OR: Odds ratio

OSH: Occupational Safety and Health

OSHA: Occupational Safety and Health Administration

PEL: Permissible Exposure Limit

PNOR: Particles not otherwise regulated

PPE: Personal protective equipment

RIE: Relative index of exposure

RR: Risk ratio

SE: Standard error

SIC: Standard Industrial Classification

STEL: Short-term exposure level

TLV: Threshold Limit Value

TWA: Time-weighted average

Remerciements

Je tiens tout d'abord à remercier mon directeur de recherche Jérôme Lavoué pour son accueil, sa confiance et son support. C'est grâce à son partage de connaissances et d'expertises, sa rigueur exemplaire et sa capacité à encourager la discussion et la réflexion que j'ai pu acquérir les compétences et capacités de recherche que je possède aujourd'hui. Il m'a également fourni les opportunités de recherche nécessaires afin que je puisse développer mon potentiel de chercheur dans le domaine de la santé au travail.

Je tiens à remercier mon employeur, l'Institut de recherche Robert-Sauvé en santé et en sécurité du travail (IRSST), pour m'avoir fourni les ressources et outils nécessaires afin que je puisse me dédier entièrement à mes études doctorales. Un merci spécial à Joseph Zayed qui a cru en mes capacités et a grandement participé à initier cette grande aventure doctorale par le soutien de ma candidature auprès de la direction de l'Institut.

J'aimerais également remercier Jean-François Sauvé, mon comparse de bureau à l'Université, qui s'est toujours montré disponible pour répondre à mes questions autant sur le domaine de l'évaluation de l'exposition que sur le (fameux…) logiciel d'analyse R.

Je tiens aussi à saluer les collaborateurs avec lesquels j'ai évolué tout au long de ces années et qui m'ont aidé dans mon cheminement: Dan Vatnik, Igor Burstyn, Laurel Kincl, France Labrèche, Robin Ackerman et Melissa Friesen.

Je tiens également à remercier mes parents, qui ont toujours eu l'éducation à cœur et qui m'ont toujours encouragé dans mes choix de carrière.

Et je termine par ce qui me semble le plus précieux; je remercie mes enfants, Marc-Antoine et Rémi, qui ont su me faire décrocher des équations et des concepts abstraits par leur joie de vivre et leur bonne humeur. Je remercie finalement du fond du cœur ma jolie femme, Judith, pour son amour, ses encouragements et pour avoir agi comme pilier de la famille dans mes moments d'absence physique (et « mentale »…) tout au long des cinq dernières années.

CHAPITRE 1- Mise en contexte

INTRODUCTION GÉNÉRALE

1.1 Introduction générale

En santé au travail, la connaissance des conditions d'exposition des travailleurs aux substances chimiques joue un rôle essentiel. La disponibilité de l'information sur l'intensité, la durée et la fréquence de l'exposition aux contaminants permet de soutenir la mise en place de programmes de surveillance de l'exposition et de politiques de prévention ciblées par secteur d'activité ou par métier. Elle est également nécessaire pour la réalisation d'études épidémiologiques et peut servir au développement de modèles prédictifs de l'exposition en milieu de travail.

L'évaluation de l'exposition des travailleurs à des contaminants chimiques demeure un élément limitant très marqué dans toute activité visant à identifier les situations les plus à risque et à prévenir le développement de maladies professionnelles (Nieuwenhuijsen, 2003). En effet, il est reconnu que les expositions subies par une personne varient dans le temps et dans l'espace, et il n'est pas inhabituel de constater d'importantes différences entre deux mesures d'exposition effectuées à des moments différents. À titre d'exemple, quantité d'études réalisées en milieu de travail ont permis d'observer ces variations de concentrations chez un travailleur pendant la journée ou d'une journée à l'autre. Conséquemment, l'évaluation de l'exposition ne consiste pas à estimer une valeur unique au moyen de plusieurs essais (ex : mesure du point d'ébullition d'une substance), mais cherche plutôt à décrire une population de valeurs, nécessitant ainsi une grande quantité de mesures. En raison des coûts importants associés à la mesure directe de l'exposition professionnelle pour l'ensemble des circonstances rencontrées en milieu de travail, et de l'impossibilité de mesurer directement l'exposition passée, les données historiques préexistantes constituent une source précieuse d'information.

1.1.1 Les banques de données d'exposition professionnelle (BDEP)

Les BDEP, mises en place au début des années 1980 par plusieurs pays européens et nord-américains, représentent une source majeure de mesures d'exposition historiques couvrant de multiples activités industrielles et agresseurs chimiques. Ces banques de données contiennent les mesures d'exposition recueillies par les agences gouvernementales dans le cadre d'activités de prévention ou de contrôle de respect des normes, les concentrations mesurées étant associées à un certain nombre de variables les caractérisant (par exemple : secteur d'activité, raison de la visite, type d'exposition). Les pays pour lesquels ce type de BDEP a été décrit dans la littérature incluent la France (Vincent et Jeandel, 2001; Mater et coll., 2016), l'Allemagne (Gabriel, 2006; Koppisch et coll., 2012), le Royaume-Uni (Burns et Beaumont, 1989), l'Italie (Scarselli et coll., 2007), la Norvège (Lenvik et coll., 1999), la Finlande (Kauppinen, 2001), Singapour (Tang et coll., 2006), les États-Unis (Stewart et Rice, 1990), le Canada (Hall et coll., 2014) et le Québec (Lavoué et coll., 2012). Le tableau suivant présente quelques caractéristiques des principales BDEP:

Table I : Principales banques de données d'exposition professionnelle

Pays	Nom	Année de création	Propriétaire du système	Nombre de mesures	Objectif principal de la mise en place
États-Unis	IMIS	1979	OSHA[1]	1 500 000	▪ Vérifier la conformité aux normes réglementaires
France	COLCHIC	1987	INRS[2]	800 000	▪ Prévention
Québec	LIMS	1984	IRSST[3]	900 000	▪ Prévention ▪ Vérifier la

					conformité aux normes réglementaires
Allemagne	MEGA	1972	BG[4]	2 200 000	▪ Prévention ▪ Assurance
Royaume-Uni	NEDB	1986	HSE[5]	150 000	▪ Prévention ▪ Élaboration de politiques réglementaires
Italie	SIREP	1996	ISPESL[6]	100 000	▪ Prévention ▪ Réduction du risque cancérogène

[1] Occupational Safety and Health Administration
[2] Institut National de Recherche et de Sécurité
[3] Institut de recherche Robert-Sauvé en santé et en sécurité du travail
[4] BG Institut d'Hygiène et de Sécurité
[5] Health and Safety Executive
[6] Institut Italien pour la Sécurité et la Prévention

Les mesures contenues dans les BDEP sont actuellement exploitées pour évaluer l'exposition dans plusieurs efforts de recherche en santé au travail. Une vaste étude internationale actuelle coordonnée par le Centre International de Recherche sur le Cancer (CIRC), appelée SYNERGY, vise à caractériser l'association entre des facteurs de risque professionnels et le cancer du poumon (Olsson et coll., 2011; Peters et coll., 2012a; Bigert et coll., 2015). Dans le cadre de ce projet, une banque de données contenant plus de 350 000 mesures d'exposition au nickel, chrome hexavalent, hydrocarbures aromatiques polycycliques (HAP), silice cristalline et amiante a été assemblée en majeure partie à partir des banques COLCHIC, MEGA et NEDB (Peters et coll., 2011; Peters et coll., 2012b). Un autre effort de recherche actuel vise quant à lui à développer un outil prédictif de l'exposition nommé Advanced REACH Tool (ART) (Tielemans et coll., 2007; Tielemans et coll., 2008; Tielemans et coll., 2011; van Tongeren et coll., 2011; Schinkel et coll., 2013; McNally et coll., 2014). ART est un outil générique permettant de combiner les prédictions générées par un modèle physique déterministe d'exposition avec des données d'exposition disponibles, provenant notamment

des banques MEGA et NEDB, via une approche Bayésienne. De son côté, l'Institut National de Recherche et de Sécurité (INRS) a produit deux outils en ligne à partir des données quantitatives de la banque française COLCHIC fournissant de l'information sur l'exposition professionnelle aux fibres (INRS, 2015a) et aux solvants (INRS, 2015b). Notons également des analyses réalisées avec la banque italienne SIREP ayant permis de dresser des portraits globaux de l'exposition professionnelle à des agents cancérogènes tels le chrome hexavalent, le benzène et les poussières de bois (Scarselli et coll., 2011; Scarselli et coll., 2012; Scarselli et coll., 2013). Ces exemples de résultats permettent d'apprécier le potentiel des BDEP en tant que source d'information pour la réalisation de différentes activités en santé au travail.

1.1.2 La banque IMIS

Accessible publiquement, la BDEP américaine IMIS (Integrated Management Information System) est la plus importante source multi-industries d'information sur l'exposition en Amérique du Nord. Elle a été mise en place en 1979 par l'agence fédérale américaine OSHA et contient aujourd'hui près de 1,5 million de mesures de l'exposition effectuées par les inspecteurs d'OSHA chargé de vérifier la conformité aux valeurs limites d'exposition (VLE). Les mesures quantitatives d'exposition sont accompagnées d'informations en lien avec l'inspection telles que les caractéristiques de l'industrie visitée, le secteur d'activité, la date de prélèvement, la raison de la visite, le type d'exposition et le titre d'emploi échantillonné. La banque IMIS présente donc le potentiel le plus élevé pour le développement d'applications liées à la prévention dans le contexte nord-américain. Parmi les utilisations récentes de cette banque, des analyses ont été présentées pour le plomb (Okun et coll., 2004; Henn et coll., 2011), la silice cristalline (Linch et coll., 1998; Yassin et coll., 2005), le formaldéhyde (Melville et Lippmann, 2001; Lavoue et coll., 2008), le béryllium (Hamm et

Burstyn, 2011), l'amiante (Cowan et coll., 2015) et les hydrocarbures aromatiques polycycliques (Lee et coll., 2015): ces études ont permis de dresser des portraits globaux de l'exposition professionnelle, d'estimer la quantité de travailleurs exposés dans certains secteurs industriels et d'établir l'évolution temporelle de l'exposition à ces contaminants d'intérêt.

En complément aux données IMIS, l'agence OSHA a rendu disponible sur son site internet en 2010 une banque de résultats d'analyse nommée Chemical Exposure Health Data (CEHD). Cette banque contient les résultats analytiques de laboratoire correspondant aux prélèvements effectués par les inspecteurs d'OSHA lors de leurs visites de conformité (Lavoue et coll., 2013; OSHA, 2015). Toutefois, la banque CEHD ne chevauche que partiellement la banque IMIS puisque les enregistrements dans CEHD sont limités aux échantillons recueillis par les agences assujetties au régime fédéral (federal OSHA). Aux États-Unis, l'encadrement réglementaire est assuré par le régime fédéral d'OSHA dans tous les États, à l'exception de ceux qui ont adopté une législature d'État en ce sens (state OSHA plan). En date de 2016, 21 États assuraient la mise en œuvre de la législation en matière de prévention en santé et sécurité au travail (SST) (OSHA, 2015). Le laboratoire d'analyse de Salt Lake City, créé en 1984, traite l'ensemble des échantillons recueillis par les agences assujetties au régime fédéral.

1.1.3 Représentativité des mesures enregistrées dans les BDEP - principe

Bien que les BDEP représentent un très fort potentiel d'information quantitative sur l'exposition, de nombreux utilisateurs de ces banques ont mentionné la présence potentielle d'une différence systématique entre les niveaux rapportés dans une BDEP et les niveaux

d'exposition dans la population de travailleurs (c.-à-d. biais). Ainsi, les niveaux d'exposition obtenus à partir des visites d'inspection ou de prévention pourraient théoriquement surestimer ou sous-estimer l'exposition des travailleurs non échantillonnés.

Les banques d'informations sur l'exposition professionnelle créées à partir des mesures recueillies lors des visites en industrie ne peuvent donc être interprétées directement comme un portrait représentatif des conditions de travail en général. Ces mesures ne résultent pas d'un plan d'échantillonnage aléatoire des milieux de travail. L'ensemble des étapes menant à la prise de mesure puis à l'enregistrement d'un niveau d'exposition dans une banque de données peuvent introduire un biais (Olsen et coll., 1991). La figure suivante présente un schéma conceptuel illustrant ces étapes :

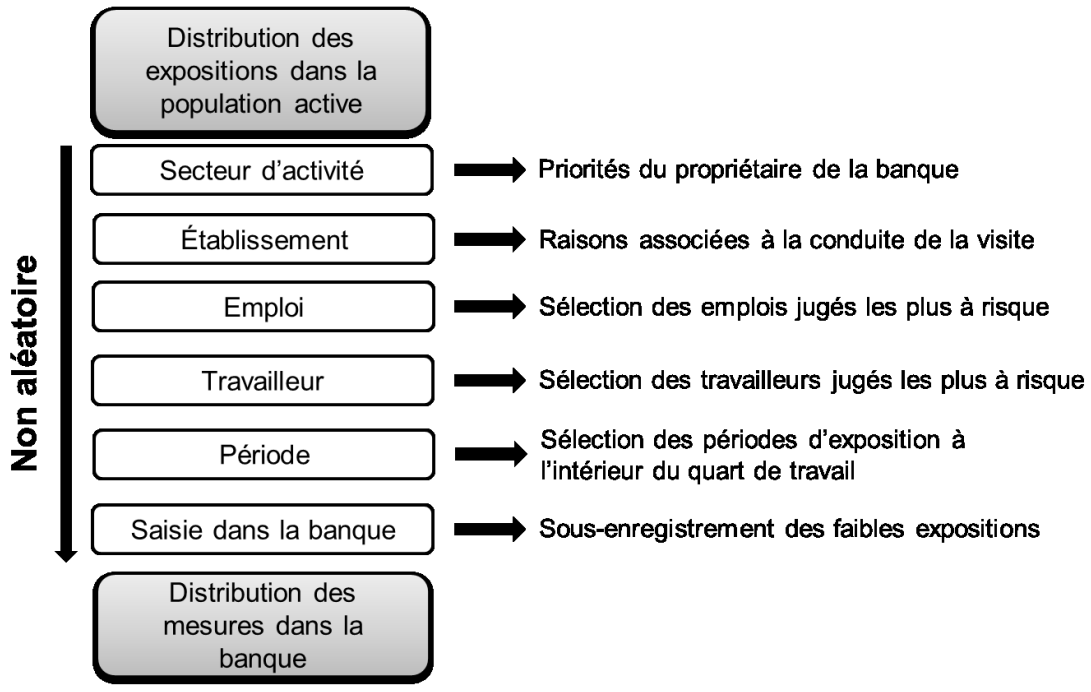


Figure 1 : Schéma conceptuel des biais dans une banque de données d'exposition professionnelle (adapté de Lavoué (2006))

Traditionnellement, dans le but d'optimiser l'utilisation des ressources disponibles, les hygiénistes ont visé à prendre des mesures pendant des périodes ou des circonstances associées à de fortes expositions. Cette stratégie d'échantillonnage causerait ainsi une surreprésentation de valeurs élevées et donc une surévaluation des niveaux réels. Il est également possible que les intervenants jugent non-pertinent d'enregistrer les faibles expositions dans la banque de données, telles que les valeurs non décelées ou nettement sous la norme, ce qui causerait également une surévaluation des niveaux réels. En revanche, certaines stratégies d'échantillonnage sont susceptibles de mener à une sous-évaluation des niveaux réels. À titre d'exemple, il est possible que des mesures soient prises sur une tâche lors d'inspections faisant suite à une plainte de travailleur même si l'intervenant juge que l'exposition est nulle (p. ex. en raison d'une obligation réglementaire), ce qui pourrait mener à une surreprésentation de faibles valeurs.

En théorie, la meilleure façon d'étudier les biais associés aux mesures contenues dans les BDEP serait de comparer les portraits établis à partir de ces mesures à des échantillons aléatoires provenant de l'ensemble des milieux de travail couverts par les intervenants. Cette approche directe ne peut être utilisée puisqu'on ne dispose presque jamais de mesures résultant de campagnes de mesures aléatoires. L'étude des biais a donc historiquement été abordée de façon indirecte selon deux angles : la première approche consiste à étudier l'association entre les niveaux d'exposition enregistrés dans la banque et des éléments contextuels (variables internes) potentiellement associés à la stratégie de mesure alors que la deuxième approche consiste à comparer le contenu de la banque avec des sources de données externes.

Association entre les niveaux d'exposition enregistrés dans une BDEP et les variables internes:

Les BDEP contiennent, au-delà des mesures quantitatives d'exposition, plusieurs éléments d'information contextuelle reliés directement ou indirectement à la stratégie d'échantillonnage. Cette approche consiste à examiner et à estimer quantitativement l'association entre ces variables et les niveaux d'exposition. Les termes « déterminant » et « prédicteur » utilisés dans la littérature (et dans cette thèse) pour décrire l'association entre une variable et les niveaux d'exposition n'impliquent pas de lien de causalité.

Les deux exemples suivants permettront d'illustrer le type de résultat pouvant être obtenu pour les variables « taille de l'entreprise » et « raison de la visite » et les implications sur l'interprétation des estimés effectués à partir d'une BDEP.

1) Variable « taille de l'entreprise » : s'il est démontré que les niveaux d'exposition sont systématiquement plus faibles lors des évaluations effectuées dans les entreprises de grande taille, il peut s'avérer nécessaire d'ajuster les prédictions faites à partir du contenu de la BDEP. En effet, si l'on observe dans un secteur d'activité particulier que la proportion d'entreprises de grande taille est supérieure dans cette BDEP par rapport à la population, le fait de ne pas considérer cette différence de distribution causera une sous-estimation des niveaux d'exposition réellement rencontrés dans ce secteur d'activité. Puisqu'il est possible de connaître la distribution de ce type de variable dans la population, il serait alors possible dans un tel cas de pondérer les estimés d'exposition effectués à partir de la BDEP en fonction de la distribution des tailles d'entreprises dans la population.

2) Variable « raison de la visite » : l'observation de niveaux d'exposition systématiquement plus élevés lors des inspections faisant suite à une plainte d'employé que lors des visites programmées suggérerait un biais de sélection (les deux types de raison conduisent à la mesure de situations d'exposition particulière). Il serait encore une fois nécessaire d'effectuer un ajustement aux prédictions faites à partir de la BDEP. Cependant, il n'est pas possible pour ce type de variable de connaitre dans quelle mesure l'une ou l'autre des situations reflète réellement la population. Il faudra porter un jugement quant à laquelle des deux situations est la plus représentative de l'ensemble des milieux de travail, voire décider d'un poids à accorder aux deux catégories de la variable pour effectuer une correction des estimations. Par exemple, on pourrait simplement considérer les données « programmées » comme étant représentatives (poids = 100%) et ne pas prendre en compte les données « plainte » (poids = 0%) de nos estimés d'exposition.

Comparaison avec des sources de données externes :

La comparaison d'une BDEP avec des sources de données externes peut également représenter une source d'information sur d'éventuels biais de stratégie de mesures. L'observation de différences entre les niveaux d'exposition seulement pour certains contaminants spécifiques ou dans certains secteurs d'activité pourra s'expliquer par des particularités locales telles que la réglementation (ex : normes à portée réglementaire vs. valeurs guides) ou par des différences de procédés, de méthodes de travail ou de moyens de contrôle entre les pays. Par contre, l'identification de différences systématiques à travers l'ensemble du spectre industriel et à travers de multiples substances sera indicatrice d'un effet de stratégie globale d'échantillonnage. L'observation de similitudes globales entre une BDEP et une source de données externes permet quant à elle de renforcer la fiabilité dans les estimés d'exposition réalisés à partir de cette banque. En effet, il est peu probable que deux

sources indépendantes possédant des stratégies de sélection des milieux de travail et des situations échantillonnées très différentes puissent comporter les mêmes biais. La comparaison de sources de mesures représente toutefois un défi particulier puisque les systèmes de nomenclature utilisés varient d'une source de mesures à l'autre, en particulier au niveau des systèmes de classification des emplois et de la catégorisation des produits chimiques. À titre d'exemple, Lavoue et coll. (2011) ont dû éliminer 24% des données dans leur analyse comparative des banques IMIS et COLCHIC en raison de systèmes de classification des emplois différents utilisés par les deux banques.

1.1.4 Représentativité des mesures enregistrées dans les BDEP – résultats disponibles

Parmi l'ensemble des études retrouvées dans la littérature portant sur des biais identifiables par l'étude de paramètres internes aux BDEP, seule une étude portait sur une autre banque qu'IMIS. Peters et coll. (2011) ont montré dans cette étude que les niveaux d'exposition à la silice cristalline étaient plus élevés pour la stratégie de mesure « pire des cas » (worst-case) comparé à une stratégie de mesure représentative dans la banque de données ExpoSYN créée dans le cadre du projet SYNERGY (ratio des médianes des niveaux = 1,8). Les études basées sur IMIS ont quant à elles porté principalement sur l'influence des raisons des visites d'entreprise (planifiées, causées par des plaintes d'employé), de la taille de l'établissement et du statut syndical sur les niveaux d'exposition à un contaminant (principalement le plomb, la silice et le formaldéhyde). Henn et coll. (2011) ont notamment montré que les établissements de petite taille ainsi que les établissements syndiqués avaient une probabilité plus élevée d'avoir des résultats d'exposition dépassant la VLE d'OSHA pour le plomb. Melville et Lippmann (2001) ont quant à eux montré que les niveaux d'exposition au toluène et au formaldéhyde étaient plus élevés lors des visites causées par des plaintes d'employés

comparé aux visites planifiées. Une revue de littérature par Lavoue et coll. (2013), décrivant l'ensemble des études portant sur IMIS, a conclu que les résultats disponibles n'avaient pas montré la présence d'associations notables de façon cohérente. Ces études rapportaient des résultats contradictoires sur la direction des effets des variables et l'amplitude de ces effets était modérée.

Quatre études portant sur la comparaison de BDEP (autre qu'IMIS) avec des données externes ont été retrouvées dans la littérature. Olsen et coll. (1991) ont comparé les résultats de la banque danoise ATABAS avec un ensemble de mesures d'exposition aux solvants recueillies dans un échantillon aléatoire de compagnies du secteur industriel du meuble au Danemark. Les résultats de cette recherche ont démontré que les niveaux d'exposition au toluène étaient plus élevés dans ATABAS (médianes des mesures d'exposition plus élevée d'un facteur 5 à 10). Vinzents et coll. (1995) ont comparé les mesures de xylène dans les secteurs industriels du travail du bois et de la peinture par pulvérisation dans 5 BDEP européennes : ATABAS (Danemark), MEGA (Allemagne), COLCHIC (France), EXPO (Norvège) et NEDB (Royaume-Uni). Les résultats de cette recherche ont démontré que les niveaux d'exposition dans les deux banques constituées de données prises dans le cadre d'assurances (MEGA et COLCHIC) étaient plus faibles que dans les trois autres banques, constituées de données de conformité à des normes environnementales (ratio des médianes des niveaux = 0,25). Peters et coll. (2011) ont comparés les mesures de silice cristalline dans plusieurs BDEP européennes et une BDEP canadienne. Les résultats de cette recherche ont montré que les niveaux d'exposition étaient plus élevés dans les banques britanniques et canadiennes, alors qu'ils étaient moins élevés dans les banques allemandes et du nord de l'Europe (ratio maximal des médianes de 4,5 entre les pays). Finalement, Mater et coll. (2016) ont comparé les résultats de deux BDEP françaises pour la période 2007-2012. Les

résultats ont montré des niveaux d'exposition plus élevés dans la banque de prévention COLCHIC comparé à la banque de conformité réglementaire SCOLA (ratio des médianes des niveaux = 3,45).

Cinq études portant sur la comparaison de la banque IMIS avec des données externes ont été retrouvées dans la littérature. Okun et coll. (2004) ont comparé, pour le plomb, le contenu de la banque IMIS (banque de « conformité réglementaire ») avec une banque intégrant des mesures associées à une stratégie préventive (banque de « prévention »). Les résultats de cette recherche ont démontré que les niveaux d'exposition associés à une stratégie de conformité étaient systématiquement plus élevés que ceux étant liés à une stratégie de prévention, cette différence étant toutefois assez faible (proportion de mesures dépassant la PEL plus élevée de 5% dans IMIS). Lavoue et coll. (2011) ont comparé les niveaux d'exposition au formaldéhyde entre IMIS et la banque française COLCHIC. Malgré des milieux de travail et des profils industriels potentiellement très différents entre les deux pays, les portraits globaux de l'exposition au formaldéhyde établis à partir des banques américaines et françaises étaient similaires. Finalement, Mendeloff (1984), Jones et coll. (1986) et Lavoue et coll. (2013) ont comparé le contenu de la banque IMIS avec la banque de résultats d'analyse CEHD pour différents agents chimiques. Ces recherches ont montré que les résultats des mesures effectuées par les inspecteurs d'OSHA n'étaient pas systématiquement enregistrés dans la banque IMIS. Ces études rapportaient cependant des résultats contradictoires sur l'existence d'un sous-enregistrement plus marqué pour les résultats faibles ou non détectés.

1.1.5 Les mesures non-détectées dans IMIS

L'analyse des niveaux d'exposition est rendue complexe par la présence d'une proportion élevée de résultats non décelés (ND). Plusieurs auteurs se sont intéressés à l'interprétation des résultats ND dans IMIS depuis son instauration au début des années 1980 (Froines et coll., 1986; Froines, 1989; Teschke et coll., 1999; Melville et Lippmann, 2001; Henneberger et coll., 2004; Henn et coll., 2011; Lavoue et coll., 2013) et ont discuté de trois défis particuliers. Premièrement, il a été suggéré qu'une proportion inconnue de résultats ND correspondaient à des situations de « non-présence » (c.-à-d. agent était absent du milieu de travail), le reste des résultats ND correspondant à des situations où l'exposition était « présente mais non détectée » par la méthode analytique. Des auteurs ont mentionné que la quantité élevée de résultats ND dans IMIS suggérait qu'une forte proportion de ND pourrait correspondre à l'interprétation de « non-présence » (Melville et Lippmann, 2001; Henn et coll., 2011; Lavoue et coll., 2013). Par contre, il existe présentement peu de données empiriques sur cette problématique, et ce phénomène pourrait se révéler spécifique au contexte dans lequel la mesure a été recueillie (p. ex. secteur d'activité particulier). Deuxièmement, aucune information sur la valeur de censure (limite de quantification) n'est disponible dans IMIS lorsqu'un enregistrement est rapporté comme ND. Troisièmement, tel que discuté dans plusieurs études (Hamm et Burstyn, 2011; Lavoue et coll., 2013; Lee et coll., 2015), le statut d'une mesure codée ND est indiqué dans la même variable identifiant si l'échantillon correspond à une valeur d'exposition moyenne pondérée sur 8 heures (VEMP-8h) ou une valeur d'exposition de courte durée (VECD). Ceci a pour conséquence d'empêcher un utilisateur de savoir si un résultat ND correspond à une mesure VEMP-8h ou une mesure VECD (correspondant en général à une limite de quantification plus élevée).

L'absence d'information sur la valeur de censure dans IMIS complique l'interprétation et le traitement des résultats ND. Puisque cette valeur est dépendante de l'année (amélioration de

la sensibilité des techniques analytiques) et de la durée de la mesure (volume d'air prélevé plus important), la signification d'un résultat ND pour un agent chimique pourrait refléter des expositions potentiellement différentes. Il est effectivement possible que la limite de quantification pour un résultat ND soit très différente de celle d'un autre résultat ND, p. ex. pour une mesure court terme en 1979 (LOQ plus élevée) et une mesure long terme en 2011 (LOQ plus faible). En théorie, l'amélioration des techniques au cours des dernières décennies devrait avoir contribué à diminuer la quantité d'échantillons mesurés sous la limite de détection. Okun et coll. (2004) ont toutefois observé une augmentation de la proportion de résultats ND dans IMIS pour le plomb entre 1979 et 1997 (de 20% à 50%). Ceci suggère que d'autres facteurs (p. ex. tendances temporelles liées à l'exposition à des contaminants en milieu de travail généralement à la baisse (Symanski et coll., 1998; Creely et coll., 2007)) ont pu quant à eux contribuer à augmenter le nombre de résultats ND à travers les années. Un projet visant à attribuer la limite de quantification correspondante à chaque résultat ND présent dans IMIS (en tenant compte de l'année et de la durée de mesure) est actuellement en développement dans l'équipe du Pr Lavoué, ce qui devrait faciliter à terme le traitement de ces résultats.

L'approche la plus courante dans le domaine de l'hygiène du travail permettant de traiter les résultats ND consiste à substituer ces valeurs par la limite de quantification divisée par un facteur 2 ou la racine carrée de 2 (Hornung et Reed, 1990). Des méthodes considérées plus valides sont également disponibles, telles que la régression de statistiques d'ordre, l'estimateur du maximum de vraisemblance et la méthode Kaplan-Meier (Lubin et coll., 2004; Hewett et Ganser, 2007). Considérant le peu d'information entourant les mesures ND dans IMIS, ces résultats ont généralement été exclus des analyses dans les études antérieures (Froines et coll., 1990; Melville et Lippmann, 2001; Lurie et Wolfe, 2002), ou plus

récemment, inclus en les classant comme inférieur à un seuil d'exposition prédéfini largement supérieur à la limite de détection, par exemple une valeur limite d'exposition (Hamm et Burstyn, 2011; Henn et coll., 2011). Pour cette dernière approche, la réponse n'est plus la concentration mesurée mais plutôt une variable indiquant si cette concentration est inférieure ou supérieure au seuil choisi : la limite de détection n'est donc plus importante pourvu qu'elle soit inférieure à ce seuil.

Deux études ayant évalué l'impact de l'exclusion des résultats ND dans IMIS ont été retrouvées dans la littérature (Lavoue et coll., 2008; Lavoue et coll., 2011). Différents scénarios de distribution des ND dans les catégories VEMP-8h et VECD ont été simulés dans ces études pour le formaldéhyde (c.-à-d. en les distribuant dans les catégories VEMP-8h ou VECD selon leurs proportions respectives dans les résultats détectés ou selon des proportions prédéfinies). Ces études ont conclu à des impacts non-négligeables sur les niveaux d'exposition prédits.

1.1.6 Limite des études existantes

En résumé, les données quantitatives contenues dans les BDEP représentent une source d'information importante sur les conditions d'exposition des travailleurs aux substances chimiques. La BDEP américaine IMIS est celle qui présente le potentiel le plus élevé pour l'amélioration des efforts de prévention des maladies professionnelles dans le contexte nord-américain. Cependant, la représentativité des données IMIS par rapport à la distribution réelle des niveaux d'exposition retrouvés dans la population générale est largement inconnue. Certains efforts de recherche se sont intéressés à identifier et caractériser les biais potentiels présents dans la banque IMIS depuis sa mise en place il y a une trentaine d'année, mais les

résultats obtenus ne sont que parcellaires. Le nombre d'études est limité, ces travaux ayant majoritairement porté sur une seule substance à la fois et sur des sous-ensembles particuliers de données. Les protocoles d'analyse utilisés différaient d'une étude à l'autre, ce qui complique également l'interprétation des résultats. Il importe donc de mettre en place le premier effort systématique d'évaluation des données d'hygiène contenues dans la banque IMIS, portant sur l'ensemble des agents chimiques et couvrant tout le spectre industriel et l'entière période temporelle, ce qui permettra de mieux caractériser les biais potentiels et d'améliorer l'interprétation des résultats disponibles. De plus, la disponibilité récente de la banque CEHD ouvre des perspectives additionnelles venant de la comparaison des deux banques, en particulier l'étude du sous-enregistrement et de la durée d'échantillonnage des résultats ND.

1.2 Objectifs de la recherche

1.2.1 Objectif général

L'objectif général de cette recherche est d'établir dans quelle mesure les données de contamination de l'air recueillies par l'agence fédérale américaine OSHA peuvent être utilisées pour l'estimation des expositions professionnelles en Amérique du Nord

1.2.2 Objectifs spécifiques de la recherche

Le premier volet de cette recherche vise à étudier à travers l'ensemble des agents chimiques si les variables reflétant les caractéristiques des établissements visités et des inspections sont associées avec les niveaux d'exposition contenus dans IMIS.

Le second volet vise à étudier à travers l'ensemble des agents chimiques les facteurs déterminants l'enregistrement des échantillons de CEHD dans la banque IMIS.

Le troisième volet vise à prédire, parmi les résultats non détectés (ND) de la banque IMIS, lesquels correspondent à des valeurs d'exposition moyenne pondérée sur 8 heures (VEMP-8h) ou à des valeurs d'exposition de courte durée (VECD) en se basant sur l'information contenue dans la banque CEHD.

1.3 Organisation de la thèse

Cette thèse, divisée en six chapitres, présente au chapitre 1 la problématique de l'utilisation des banques de données d'exposition professionnelle (BDEP) en général et de la banque IMIS en particulier pour l'évaluation de l'exposition professionnelle. Le chapitre 2 décrit les méthodes utilisées dont la sélection des données et des variables dans les banques IMIS et CEHD, les techniques de modélisation statistiques ainsi que la méthode de méta-analyse utilisée pour synthétiser les résultats obtenus à travers les agents. Les trois chapitres suivants contiennent les manuscrits d'article scientifique constituant la contribution principale de cette thèse. Le chapitre 6 présente la discussion générale des résultats et les conclusions tirées de ce travail.

ARTICLE 1: Trends in OSHA compliance monitoring data 1979-2011: statistical modeling of ancillary information across 77 chemicals. Publié dans *Annals of Occupational Hygiene* (2016) DOI: 10.1093/annhyg/mev092.

ARTICLE 2: Characterization of the selective recording of sample results in OSHA's IMIS databank, 1984-2009: statistical modeling of ancillary information across 78 chemicals. Soumis aux co-auteurs en vue d'une publication dans une revue scientifique.

ARTICLE 3: Non-detects in OSHA's IMIS databank, are they short term or 8-hour shift-long samples? Prediction for 54 chemicals using recursive partitioning statistical methods. Soumis aux co-auteurs en vue d'une publication dans une revue scientifique.

CHAPITRE 2- Méthodologie

2.1 Description de la banque de données d'exposition IMIS

La banque de données IMIS de l'agence américaine OSHA contient l'ensemble des informations historiques sur les visites d'inspection effectuées par les inspecteurs d'OSHA chargés de vérifier la conformité aux valeurs limites d'exposition (VLE). Les mesures quantitatives recueillies par OSHA représentent un sous-ensemble spécifique de la banque IMIS. Un extrait électronique de la banque IMIS, couvrant la période 1979-2011, a été obtenu par l'équipe de recherche du Professeur Jérôme Lavoué auprès de l'agence fédérale américaine OSHA via la loi américaine sur la liberté d'accès à l'information (Freedom of Information Act).

Brièvement, l'extrait IMIS contenait 851 987 enregistrements associés à 132 280 inspections effectuées en milieu de travail pour la période 1979-2011, ces données étant réparties entre 1 050 codes d'activité industrielle selon la classification américaine Standard Industrial Classification (SIC) de 1987 (OSHA, 2014). Les mesures sont accompagnées d'informations contextuelles permettant de connaître les caractéristiques de l'inspection (Tableau I). Étant donné que les données d'IMIS proviennent d'activités de vérification de la conformité aux normes, leur majorité consiste en des moyennes pondérées sur 8 heures (VEMP-8h) ou encore en des moyennes de courte durée (VECD) en fonction de la période de référence spécifiée par OSHA. Des 1 169 différents agents chimiques présents dans la banque, 65 avaient fait l'objet de plus de 1 000 mesures d'exposition personnelle ou d'ambiance, représentant 91% du nombre total d'enregistrements dans IMIS. Ces agents étaient majoritairement répartis dans les classes d'agents chimiques suivantes: métaux, solvants, gaz, poussières, isocyanates et HAP. Finalement, un peu plus de 65% des données

correspondaient à des inspections faisant suite à une plainte d'employé ou la référence d'un inspecteur, les autres mesures étant associées à des visites planifiées.

2.2 Description de la banque de résultats d'analyse CEHD

En complément aux données IMIS, l'agence OSHA a rendu disponible sur son site internet en 2010 la banque CEHD contenant les résultats analytiques de laboratoire correspondant aux prélèvements réalisés par les inspecteurs d'OSHA lors de leurs visites de conformité effectuées depuis 1984 (Lavoue et coll., 2013; OSHA, 2015). Les données CEHD complémentent les données IMIS en fournissant de l'information sur la durée d'échantillonnage, la méthode analytique ainsi que sur la présence d'autres agents chimiques mesurés à partir du même support d'échantillonnage (Tableau I). Le personnel d'OSHA réalise des calculs à partir des résultats analytiques (p. ex. calcul d'une VEMP-8h à partir de plusieurs échantillons consécutifs de courte durée) et enregistre par la suite le résultat de leur calcul dans IMIS.

L'extrait électronique de CEHD contenait 1 450 836 enregistrements pour la période 1984-2009. La majorité de ces enregistrements correspondait à des mesures personnelles (78.4%), le reste étant constitué de mesures en postes stationnaires, frottis de surface et d'échantillons en vrac (Lavoue et coll., 2013). Des 1 082 différents agents chimiques présents dans la banque, 78 avaient fait l'objet de plus de 1 000 mesures d'exposition personnelle.

2.3 Liaison des banques IMIS et CEHD

Les analyses effectuées pour les articles 2 et 3 ont nécessité le croisement des données de la banque IMIS avec les données de la banque CEHD. Pour ces deux études, les deux sources

de données ont été combinées en utilisant la variable 'sampling number', considérée comme identificateur unique d'une 'évaluation' réalisée par un inspecteur (p. ex. quart de travail de 8h d'un travailleur). Il est à noter que cette variable ne pouvait toutefois être considérée comme un identificateur parfait permettant de lier IMIS et CEHD. Ainsi, la durée d'échantillonnage totale des échantillons consécutifs ayant un même 'sampling number' dans CEHD semblait irréaliste dans certains cas (p. ex. >600 minutes), et certains enregistrements dans IMIS avaient le même 'sampling number'. Cependant, des analyses préliminaires sur trois agents chimiques (plomb, formaldéhyde et toluène) ont montré que ces situations étaient relativement rares (<3%). Les multiples enregistrements liés à un 'sampling number' dans CEHD ont donc été traités comme des mesures consécutives séquentielles d'un quart de travail et agrégés pour le calcul de la durée d'échantillonnage totale et du résultat de concentration de l'évaluation.

2.4 Description des variables analysées

2.4.1 Variables présentes dans les banques IMIS et CEHD

Les variables retenues pour nos travaux ont été restreintes aux paramètres codifiés, éliminant par exemple les champs contenant des informations non standardisées tels le titre d'emploi et les commentaires généraux reliés à l'exposition. Ensuite, seules les variables dont la description était disponible pour une proportion suffisante (c.-à-d. >95%) de mesures ont été retenues.

Dans le cas de la variable correspondant à la raison ayant mené à la conduite de l'inspection ('type d'inspection'), seulement 5 des 12 catégories possibles – planifié, plainte, référé par un inspecteur, suivi et surveillance – ont été retenus pour les analyses puisqu'elles représentaient

plus de 95% des données. Les catégories 'suivi' et 'surveillance' ont de plus été combinées puisqu'elles correspondaient à des inspections où l'inspecteur devait retourner dans un établissement particulier. La variable 'taille d'établissement' a été catégorisée en séparant les données en tertiles basés sur le nombre de travailleurs dans chaque établissement inspecté (1 à 35 travailleurs = petite, 36 à 150 = moyenne, plus de 150 = grande). La variable 'type d'exposition' a été séparée en deux catégories : VEMP-8h, correspondant aux entrées 'time-weighted average (TWA)' de la banque de données IMIS, et VECD, correspondant aux entrées 'short-term exposure level (STEL)', 'peak' et 'ceiling levels' de la banque de données. Des analyses préliminaires avaient montré l'absence de différences entre les niveaux d'exposition associées à ces trois catégories de mesures de courte durée.

Table I : Sommaire des informations contenues dans les banques de données d'exposition IMIS et CEHD

Variable	Type	Description
Variables communes aux banques IMIS et CEHD		
Inspection number	Nominale	Identificateur unique lié à l'inspection
Nom de l'établissement	Texte	Nom de l'établissement lié à l'inspection (les noms ne sont pas uniques: c.-à-d. qu'il peut y avoir des variations dans la façon dont les établissements sont orthographiés)
Ville	Texte	Indique le nom de la ville où l'inspection a eu lieu
État	Nominale	Indique l'État où l'inspection a eu lieu
Code ZIP	Nominale	Indique le code ZIP où l'inspection a eu lieu
Code SIC	Nominale	Indique le code à 4 chiffres correspondant au Standard Industrial Classification Code des versions 1972 ou 1987 (les enregistrements antérieurs à 1987 sont codés selon le système de 1972)
Sampling number	Nominale	Identificateur unique lié une évaluation de l'exposition (il peut y avoir plusieurs supports d'échantillonnage (p. ex. cassettes, tubes) liés à cet identificateur, reflétant les

Variable	Type	Description
		multiples échantillons utilisés dans le calcul d'une VEMP-8h)
Année d'échantillonnage	Continue	Date à laquelle l'échantillonnage a eu lieu
Code IMIS de l'agent	Nominale	Code IMIS de la substance chimique
Nom IMIS de l'agent	Nominale	Nom de la substance chimique fourni par IMIS
Résultat d'échantillon	Continue	Résultat d'échantillon en unité de concentration
Unité de mesure	Nominale	Unité de mesure (mg/m^3, parties par million, fibres/cc)
Type d'échantillon	Nominale	Indique le type d'échantillon : personnel, poste stationnaire, sang, urine, criblage, frottis, vrac
Variables spécifiques à la banque IMIS		
Type d'inspection	Nominale	Indique la raison de la visite : les inspections sont catégorisées en visite non programmée (plainte, suivi, référence d'un inspecteur) ou programmée (planifiée, reliée à une autre inspection)
Portée de l'inspection	Nominale	Indique s'il s'agit d'une visite partielle ou complète de l'établissement
Taille de l'établissement	Continue	Indique le nombre d'employés dans l'établissement visité
Présence d'un syndicat	Nominale	Indique le statut syndical de l'établissement visité
Plan OSHA	Nominale	Activité reliée au régime fédéral OSHA (federal OSHA) ou à un régime étatique OSHA (state OSHA plan)
Type d'exposition	Nominale	Indique si la mesure correspond à une valeur d'exposition moyenne pondérée (VEMP-8h), à une valeur d'exposition courte durée (VECD) ou à un résultat non détecté
Titre d'emploi	Texte	Courte description de l'emploi
Variables spécifiques à la banque CEHD		
Type d'instrument	Texte	Description de l'instrument de laboratoire utilisé pour l'analyse
Field number	Nominale	Identificateur unique lié à un support d'échantillonnage (p. ex. cassette, tube) soumis pour analyse
Durée d'échantillonnage	Continue	Durée d'échantillonnage en minutes

Variable	Type	Description
Volume d'air échantillonné	Continue	Volume d'air échantillonné en litres

2.4.2 Variables construites à partir de la banque d'infractions d'OSHA

Ces variables ont été créées pour refléter le profil d'infraction et de pénalités infligées aux établissements. Accessible publiquement sur le site internet du Department of Labor (DOL), la banque d'infractions d'OSHA (US Department of Labor, 2014) contient l'ensemble des infractions émises aux établissements en raison du non-respect des normes en vigueur et ce, pour toutes les visites effectuées par les inspecteurs d'OSHA depuis sa création en 1970. La procédure de liaison des banques IMIS et CEHD avec la banque d'infractions d'OSHA, basée sur le numéro d'inspection et sur l'appariement des noms d'établissements, est présentée à l'Annexe 1.

Une première variable a été créée représentant l'indice historique du comportement de non-conformité d'un établissement donné en additionnant les amendes reçues au cours de l'ensemble des visites effectuées. Une série de variables a également été créée représentant le nombre et le type d'infractions émises lors d'une inspection donnée. Pour ces variables, les infractions ont été regroupées selon l'un des trois types suivants : 'évaluation de la conformité aux normes d'exposition (PEL)', 'protection respiratoire/communication des risques' et 'autres' (infractions principalement reliées aux risques physiques, mécaniques et électriques).

2.5 Méthodes d'analyse statistique

Au cours de cette recherche doctorale, plusieurs approches d'analyse statistique avancées ont été utilisées afin de répondre adéquatement aux différentes questions de recherche. La présente section présente brièvement les techniques employées, dont certaines sont communes aux trois articles.

2.5.1 La modélisation des niveaux d'exposition dans IMIS

2.5.1.1 Modèles de régression logistique et linéaire

La régression est l'une des techniques les plus utilisées pour modéliser la relation entre les niveaux d'exposition aux contaminants chimiques et des variables contextuelles.

L'approche logistique permet d'expliquer les valeurs d'une variable réponse binaire à partir d'une combinaison de variables explicatives continues ou binaires. Dans notre étude, cette approche a été utilisée afin d'estimer l'association entre les variables explicatives et la probabilité que l'exposition dépasse un seuil d'exposition prédéfini (c.-à-d. threshold limit value (TLV) (ACGIH, 2014)). Cette approche a permis l'inclusion dans notre analyse de l'ensemble des résultats ND contenus dans la banque IMIS en les classant comme inférieur à la TLV. L'association de chaque variable explicative avec les niveaux d'exposition a été quantifiée en utilisant les mesures de rapport de cotes (RC). Pour une variable de catégorie, le RC pour une catégorie spécifique représente la cote de dépasser la norme en vigueur pour cette catégorie divisé par la cote de dépasser la norme pour la catégorie de référence de cette variable (une cote est définie par le quotient entre la probabilité de dépasser la norme et la probabilité de ne pas dépasser la norme). Pour une variable continue, le RC représente le changement de cote lorsque la variable change d'une unité.

L'approche linéaire permet d'expliquer les valeurs d'une variable réponse continue à partir d'une combinaison de variables explicatives continues ou binaires. Dans notre étude, l'approche linéaire a été utilisée afin d'estimer l'association entre les variables explicatives et le logarithme des concentrations mesurées détectées. L'effet de chaque variable explicative de type catégorie a été présenté sous la forme d'Indices relatifs d'exposition (RIE) (Lavoue et coll., 2008; Sauve et coll., 2012, 2013). Le RIE de la catégorie de référence est de 100%; une catégorie avec un RIE inférieur à 100% est donc associée à des niveaux d'exposition réduits et vice-versa. Par exemple, un RIE de 50% pour une catégorie signifie que les niveaux d'exposition pour cette catégorie sont en moyenne deux fois plus faibles que ceux dans la catégorie de référence.

Les modèles traditionnels de régression logistique et linéaire font partie de la famille des modèles linéaires généralisés (GLM) (Hosmer et Lemeshow, 2000). Dans ces modèles, la relation entre la variable réponse et une variable explicative continue est habituellement modélisée par une simple droite, ce qui peut être inapproprié. Friesen et coll. (2012) et Lavoué (2006) ont notamment montré que l'utilisation d'une droite pour décrire les relations entre les niveaux d'exposition log-transformés et l'année de la mesure ou la durée d'échantillonnage pouvait dans certains cas être inadéquat. Afin de permettre l'application de relations plus complexes entre la variable réponse et les variables explicatives continues, les modèles additifs généralisés (GAM) ont été utilisés. La régression logistique a donc été remplacée par un modèle GAM avec réponse binaire, et la régression linéaire a été remplacée par un modèle GAM avec réponse continue.

2.5.1.2 Les modèles additifs généralisés (GAM)

Les modèles GAM sont une extension des modèles GLM (Hastie et Tibshirani, 1990). Ils permettent, contrairement aux modèles logistique ou linéaire traditionnels, d'assouplir la relation entre la variable réponse et une variable explicative continue par l'utilisation de fonctions de lissage. Les avantages des GAM par rapport à d'autres méthodes de lissage sont qu'ils permettent d'utiliser une variable réponse de type continu (p. ex. concentration) ou dichotomique (p. ex. présence / absence). De plus, ils permettent d'estimer l'influence de composantes linéaires et non-linéaires au sein d'un même modèle, ainsi que l'inclusion de variables de catégorie et d'interactions entre n'importe quelles composantes du modèle.

Dans l'étude de modélisation des niveaux d'exposition dans IMIS (article #1), la seule variable continue disponible était l'année de mesure. La relation entre la variable réponse et l'année a donc été étudiée par l'approche GAM (GAM « réponse binaire » et GAM « réponse continue ») afin de permettre des tendances temporelles complexes et donc un meilleur contrôle de la confusion dans l'estimation des relations entre les niveaux d'exposition et les autres variables du modèle. À noter que l'interprétation de l'effet d'une variable explicative non lissée est réalisée de la même façon que pour un modèle logistique ou linéaire traditionnel.

Dans l'étude de l'enregistrement sélectif des données dans IMIS (article #2), ce sont des fonctions polynomiales qui ont été utilisées pour assouplir la relation entre la variable réponse et les variables explicatives continues (c.-à-d. année de mesure et durée d'échantillonnage). Cette approche a été utilisée dans cet article même si elle est considérée

moins souple que l'approche GAM puisqu'il n'était pas possible d'utiliser les GAM avec les modèles de « régression de Poisson modifiée ».

2.5.2 L'approche d'inférence multimodèle

La modélisation statistique consiste à étudier la relation entre une variable réponse et une ou plusieurs variables explicatives. La présence d'une variable dans un modèle suppose qu'elle a une influence sur la réponse, alors que les variables absentes du modèle final sont considérées comme n'ayant aucun effet sur la réponse. La stratégie utilisée pour sélectionner les variables à inclure dans le modèle final a donc un impact majeur sur les conclusions pouvant être tirées d'une analyse. La méthode traditionnellement utilisée pour le développement de modèles (stepwise regression) consiste à l'ajout ou au retrait des variables selon un critère prédéfini (p. ex. valeur p) jusqu'à l'obtention d'un modèle final contenant seulement celles supposées avoir une influence sur la réponse. Cette approche est problématique puisque l'incertitude quant à la sélection du modèle final n'est pas prise en compte. De plus, le même jeu de données est utilisé pour choisir le modèle final et estimer les coefficients, ce qui tend à donner des résultats reflétant les données dont on dispose mais n'étant pas robuste à la validation externe (c.-à-d. ne reflétant pas la population d'intérêt) (Lavoue et Droz, 2009).

La procédure d'inférence multimodèle a été utilisée pour estimer les coefficients et leur variabilité plutôt qu'une méthode basée sur des tests d'hypothèse (utilisant des valeurs P), telle que la procédure stepwise. Bien qu'amplement utilisée dans les domaines de recherche appliquée, l'approche stepwise est problématique puisqu'elle est susceptible de mener à de fausses relations et à sous-estimer l'incertitude (Harrel, 2001; Burnham et Anderson, 2002; Lavoue et Droz, 2009). Afin de pallier à l'incertitude associée au choix d'un modèle final,

Raftery et coll. (1997) ont proposé une procédure de modélisation statistique basée sur l'approche Bayésienne (Bayesian model averaging (BMA)). Burnham et Anderson (2002) ont quant à eux proposé une procédure similaire à BMA mais basée sur l'approche de la théorie de l'information, récemment utilisée pour modéliser l'exposition à différents contaminants (Lavoue et Droz, 2009; Lavoue et coll., 2011; Sauve et coll., 2012, 2013). Ce type de procédure repose sur la définition a priori d'un groupe de modèles plausibles, construits à partir de combinaisons uniques des variables à l'étude. Ainsi, contrairement aux approches traditionnelles, l'inférence est effectuée à partir de tous les modèles faisant partie de la liste établie initialement. Les résultats finaux sont obtenus en agglomérant les résultats de tous les modèles en utilisant une pondération en fonction de la qualité de l'ajustement de chaque modèle aux données. Le poids relatif d'un modèle représente la probabilité qu'un modèle candidat donné soit celui qui explique le mieux la réponse en fonction des données récoltées, et était basé dans notre analyse sur le critère d'information d'Akaike (AIC) (Burnham et Anderson, 2002). Les coefficients de régression « multimodèle » sont obtenus en calculant une moyenne des coefficients de chacun des modèles, chaque valeur individuelle étant pondérée en fonction du poids relatif de chaque modèle. L'estimation des coefficients pour une variable présente uniquement dans des modèles comportant de faibles poids relatifs va donc tendre vers zéro, un phénomène appelé « shrinkage ». Ainsi, plutôt que d'en arriver à une dichotomie « importance / aucune importance » pour une variable, l'approche multimodèle fournit un portrait plus nuancé, reflétant mieux l'incertitude associée à la sélection de modèle. Avec cette approche, le calcul de l'erreur standard d'un coefficient prend également en compte l'incertitude quant au choix du modèle. En plus de la composante « variabilité intra-modèles » (erreur moyenne du coefficient à travers les modèles), l'erreur inclut une composante « variabilité inter-modèles » (variation des estimés du coefficient à travers les modèles) (Buckland et coll., 1997). Par conséquent, les estimés de l'erreur sont

typiquement plus grands (et les intervalles de confiance plus larges) que lorsqu'ils sont obtenus avec les approches traditionnelles. Au final, l'approche multimodèle permet d'améliorer à la fois l'estimation du coefficient et de l'erreur standard, réduisant ainsi le risque de faux positif comparé aux résultats obtenus avec un modèle complet ou par l'approche stepwise. À noter que le résultat de l'approche multimodèle sera très similaire à l'approche traditionnelle (c.-à-d. stepwise) lorsqu'un modèle dans la liste a un poids très élevé. Les différences entre les deux approches apparaissent lorsque plusieurs modèles dans la liste ont des poids similaires, c'est-à-dire que les données ne permettent pas de classer clairement comme l'un meilleur que l'autre.

L'approche multimodèle a été utilisée dans l'article 1 pour quantifier les associations entre les différentes variables disponibles et les niveaux d'exposition pour chacune des approches de régression (GAM « réponse binaire » et GAM « réponse continue »).

2.5.3 L'approche de « régression de Poisson modifiée »

La régression logistique est l'approche habituellement utilisée pour étudier l'association entre des variables explicatives et une variable réponse dichotomique. Cette approche fournit directement un estimé de l'effet d'une variable sous la forme d'un rapport de cotes (RC). Le risque relatif (RR) représente toutefois la mesure d'association recherchée et correspond au risque de survenue d'un événement dans un groupe particulier divisé par le risque pour un groupe de référence. Lorsque la prévalence du caractère étudié est faible dans la population, le RC représente une bonne approximation du RR (Hosmer et Lemeshow, 2000). À l'inverse, le RC aura tendance à surestimer l'effet d'une variable donnée lorsque la prévalence du caractère étudié est élevée dans la population (>25% environ).

L'approche de régression de Poisson modifiée avec estimation robuste de la variance (Poisson regression model with a robust error variance), présentée par Zou (2004) et Greenland (2004), permet l'estimation directe du RR lorsque la variable réponse est dichotomique. Cette approche de modélisation est donc particulièrement appropriée lorsque la prévalence du caractère étudié est élevée dans la population. Lorsque la régression de Poisson est appliquée à des données binomiales, l'erreur sur l'estimé du RR sera surestimée. L'utilisation de la procédure d'estimation robuste de la variance permet d'effectuer la correction appropriée. Cette approche a récemment été appliquée en hygiène du travail pour l'évaluation de l'exposition au béryllium à partir des données IMIS (Hamm et Burstyn, 2011).

Dans le deuxième volet de la recherche doctorale, l'étude de l'enregistrement des données dans IMIS a été réalisée en utilisant l'approche de « régression de Poisson modifiée »; les échantillons de la banque CEHD ont été transformés en une variable binaire indiquant si la mesure apparaissait également dans la banque IMIS (cette variable binaire était la variable réponse). Nous avons utilisé l'approche de « régression de Poisson modifiée » car la proportion de base d'enregistrement des données dans IMIS était considérée élevée (c.-à-d. >25%), le RC ne représentant donc pas une bonne approximation du RR. L'association de chaque variable explicative avec l'enregistrement des données dans IMIS a été quantifiée en utilisant les mesures de risque relatif (RR). Pour une variable, le RR pour une catégorie spécifique représente directement la probabilité d'être enregistré dans la banque IMIS pour cette catégorie divisé par la probabilité d'être enregistré pour la catégorie de référence de cette variable.

2.5.4 L'approche de méta-analyse pour la synthèse des résultats à travers les agents

Pour la modélisation des niveaux d'exposition (GAM « réponse binaire » et GAM « réponse continue ») ainsi que pour l'étude de l'enregistrement des données dans IMIS (régression de « Poisson modifiée »), l'association entre chaque variable explicative et la variable réponse a été évaluée séparément pour chacun des agents chimiques. Pour chaque catégorie de chaque variable, une mesure d'association (p. ex. RC, RIE, RR) a donc été obtenue pour chaque agent chimique (p. ex. un RC pour la catégorie « entreprise de grande taille » pour chaque agent chimique dans l'approche GAM « réponse binaire »).

L'approche de méta-analyse est une méthode statistique qui permet de combiner les résultats individuels de plusieurs études indépendantes en un résultat commun (Borenstein et coll., 2010). Dans le contexte de la présente recherche doctorale, les « études » correspondaient aux agents chimiques. La méta-analyse à effets aléatoires avec estimateur de DerSimonian et Laird (Borenstein et coll., 2010) a donc été utilisée afin de combiner les résultats de tous les agents chimiques et d'obtenir un portrait global de l'association de chaque variable explicative avec la variable réponse. L'estimation de l'effet global de chaque catégorie de chaque variable est donc la moyenne pondérée des estimations individuelles à travers les agents chimiques. La pondération dépend de la taille d'échantillon, de l'erreur type et de la distance entre l'effet agent-spécifique et l'effet moyen global.

Pour un coefficient particulier, les résultats méta-analytique pour chaque agent chimique sont présentés de façon combinée sur un graphique appelé "forest plot". Le graphique suivant présente un exemple de forest plot pour l'estimation de l'effet d'une variable de catégorie donnée dans une régression logistique:

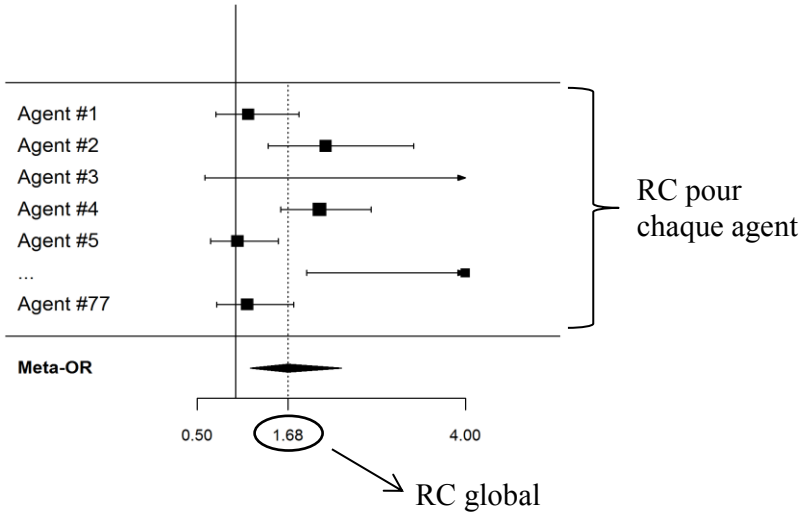


Figure 1 : RC pour chaque agent et RC global pour une variable de catégorie donnée

On retrouve sur ce graphique la description de chacun des agents chimiques, son RC individuel représenté sous forme d'un carré et son intervalle de confiance. Les résultats sont combinés sous la forme d'un losange qui représente l'effet global (c.-à-d. RC méta-analytique). Le milieu du losange représente la valeur estimée, les extrémités représentant les bornes de l'intervalle de confiance qui entourent cette estimation. La ligne verticale continue représente un effet nul de la variable (RC=1). Le forest plot permet donc de positionner les résultats de chaque agent par rapport au résultat combiné et d'ainsi présenter visuellement les différences entre agents.

2.5.5 Les modèles CART pour la prédiction de la durée de mesure des résultats ND dans IMIS

La modélisation CART (Classification And Regression Tree) est une méthode de classification basée sur la construction d'un arbre de décision binaire qui a pour but de construire des sous-groupes qui soient le plus homogène possible pour une caractéristique donnée (c.-à-d. la variable réponse). Cette approche statistique a récemment été appliquée dans différentes études en santé au travail et en épidémiologie (Friesen et coll., 2013; Wheeler et coll., 2013; Wheeler et coll., 2014; Van Hulst et coll., 2015). Brièvement, cette approche permet de construire un arbre de décision par le biais de divisions successives en fonction de variables explicatives qui peuvent être continues ou de catégorie. La variable qui prédit le mieux la séparation du jeu de données en deux sous-groupes homogènes (c.-à-d. où les données dans chaque sous-groupe sont les plus semblables en termes de la réponse étudiée) est tout d'abord identifiée. Les données sont ensuite séparées, le processus est réappliqué séparément à chacun des sous-groupes, et ainsi de suite de façon récursive jusqu'à ce que les sous-groupes atteignent une taille minimale ou que la valeur de la variable réponse soit la même pour toutes les données du sous-groupe.

Dans le troisième volet de la recherche doctorale, l'approche CART a été utilisée afin de prédire, parmi les résultats non détectés (ND) de la banque IMIS, lesquels correspondent à des valeurs d'exposition moyenne pondérée sur 8 heures (VEMP-8h) ou à des valeurs d'exposition de courte durée (VECD), en se basant sur les caractéristiques de la mesure. Le graphique suivant présente un exemple fictif simplifié d'arbre de classification :

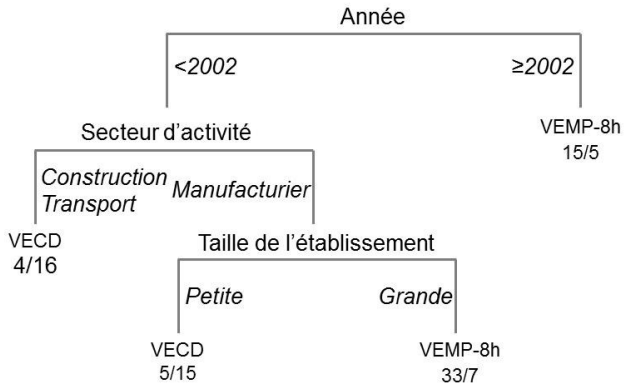


Figure 2 : Illustration d'un exemple fictif simplifié d'arbre de classification

Arbre de classification (CART) utilisant trois variables prédictives pour la classification de 100 enregistrements dans deux sous-ensembles de données. Les nœuds terminaux (terminal nodes) à la base de l'arbre présentent la distribution du nombre d'enregistrements pour chacun des sous-ensembles de données (selon l'ordre suivant : VEMP-8h / VECD) et le nom de la catégorie la plus fréquente. Dans cet exemple, le processus de modélisation prédit que les variables « année d'échantillonnage <2002 », « secteur d'activité = manufacturier » et « taille de l'établissement = grande » ont davantage de chance de mener à la catégorie « VEMP-8h » (on note que 33 des 40 enregistrements suivant cette voie correspondent à cette catégorie). La catégorie la plus fréquente est habituellement utilisée pour prédire la catégorie du nœud terminal (p. ex. VEMP-8h pour ≥2002), mais il est possible de choisir des critères différents si l'erreur est plus couteuse dans un sens que dans l'autre.

CHAPITRE 3- Trends in OSHA compliance monitoring data 1979-2011: statistical modeling of ancillary information across 77 chemicals

Article publié dans “Annals of occupational hygiene”, 2016, Vol 60 (4), pp.432-452

Trends in OSHA compliance monitoring data 1979-2011: statistical modeling of ancillary information across 77 chemicals

Philippe Sarazin (1,5), Igor Burstyn (2), Laurel Kincl (3), Jérôme Lavoué (4,5)

(1) Institut de recherche Robert-Sauvé en santé et en sécurité du travail, Chemical and Biological Hazards Prevention, Montréal, Québec, Canada

(2) Drexel University, Environmental and Occupational Health, Philadelphia, Pennsylvania, United States

(3) Oregon State University, College of Public Health and Human Sciences, Corvallis, Oregon, United States

(4) University of Montreal Hospital Research Centre, Montréal, Québec, Canada

(5) Université de Montréal, Department of Occupational and Environmental Health, Montréal, Québec, Canada

Corresponding Author:
Philippe Sarazin
Institut de recherche Robert-Sauvé en santé et en sécurité du travail
505, Boul. de Maisonneuve Ouest
Montréal, QC, Canada
H3A 3C2

Email: philippe.sarazin@irsst.qc.ca
Phone: (514) 288-1551 (Ext. 402)

Philippe Sarazin a contribué de façon majeure à la conception et au design de l’étude, a été responsable de l’analyse des données, de l’interprétation et de la présentation des résultats, de la rédaction du manuscrit et de l’édition de l’article.

3.1 Abstract

Objectives: The Integrated Management Information System (IMIS) is the largest multi-industry source of exposure measurements available in North America. However, many have suspected that the criteria through which worksites are selected for inspection are related to exposure levels. We investigated associations between exposure levels and ancillary variables in IMIS in order to understand the predictors of high exposure within an enforcement context.

Methods: We analyzed the association between 9 variables (reason for inspection, establishment size, total amount of penalty, OSHA plan, OSHA region, union status, inspection scope, year, and industry) and exposure levels in IMIS using multimodel inference for 77 agents. For each agent, we used two different types of models: (i) logistic models were used for the odds ratio (OR) of exposure being above the Threshold Limit Value (TLV) and (ii) linear models were used for exposure concentrations restricted to detected results to estimate percent increase in exposure level, i.e. relative index of exposure (RIE). Meta-analytic methods were used to combine results for each variable across agents.

Results: A total of 511,047 exposure measurements were modelled for logistic models and 299,791 for linear models. Higher exposures were measured during follow-up inspections than planned inspections (meta-OR=1.61, 95% CI: 1.44–1.81; meta-RIE=1.06, 95% CI: 1.03–1.09). Lower exposures were observed for measurements collected under state OSHA plans compared to measurements collected under federal

OSHA (meta-OR=0.82, 95% CI: 0.73–0.92; meta-RIE=0.86, 95% CI: 0.81–0.91). A 'high' total historical amount of penalty relative to none was associated with higher exposures (meta-OR=1.54, 95% CI: 1.40–1.71; meta-RIE=1.18, 95% CI: 1.13–1.23).

Conclusions: The relationships observed between exposure levels and ancillary variables across a vast majority of agents suggest that certain elements of OSHA's process of selecting worksites for inspection influence the exposure levels that OSHA inspectors encounter. Nonetheless, given the paucity of other sources of exposure data and the lack of a more demonstrably representative data source, our study considers the use of IMIS data for the estimation of exposures in the broader universe of worksites in the United States.

3.2 Introduction

Assessment of workplace exposures is an important component of the prevention and management of workplace risks. Information on intensity, duration, and frequency of exposure to chemical air contaminants is used in exposure surveillance programs (Gomez, 1997a; Ruttenber *et al.*, 2001; LaMontagne *et al.*, 2002), identification of intervention priorities (Froines *et al.*, 1986a), and epidemiologic research (Stewart and Rice, 1990; Olsson *et al.*, 2011; Friesen *et al.*, 2012; Koh *et al.*, 2012). In all of these contexts, readily available sources of measurements that represent a wide variety of chemical agents, industries, and time periods are needed. Given the significant costs associated with the direct measure of workplace exposure and the impossibility of measuring past exposures, existing exposure data can be a valuable source of information.

Nationwide occupational exposure databanks (OEDBs) are one source of individual exposure measurements (Gomez, 1993; Rajan *et al.*, 1997; Lavoue *et al.*, 2013). Established in several countries in the early 1980s, existing OEDBs contain measurements made by governmental agencies for various purposes including regulatory and prevention activities.

The largest multi-industry source of exposure measurements available in North America is the Integrated Management Information System (IMIS) maintained by the Occupational Safety and Health Administration (OSHA). IMIS has recently been replaced by the OSHA Information System (OIS) (US Department of Labor, 2014a). This

databank contains air sampling measurement results from surveys performed by OSHA inspectors to verify compliance with Permissible Exposure Limits (PELs). Several authors have reported the use of IMIS data for various purposes including the production of general portraits of exposure levels for a pollutant in an industry or an occupation (Coble *et al.*, 2001; Yassin *et al.*, 2005; Lavoue *et al.*, 2008), estimation of proportions or numbers of workers exposed in a particular industry (Linch *et al.*, 1998; Henneberger *et al.*, 2004), study of time trends in exposure (Middendorf, 2004; Okun *et al.*, 2004), identification of factors associated with exposure levels (Gomez, 1997b; Melville and Lippmann, 2001; Lavoue *et al.*, 2008; Henn *et al.*, 2011), and in support of exposure assessment in epidemiology (Hamm and Burstyn, 2011; Lee *et al.*, 2015).

Even if IMIS represents a great potential source of information, many have questioned whether the measurements reflect exposures in the general U.S. working population. The targeting of workplaces for enforcement visits and the selection of workers for exposure monitoring are non-random and may over represent situations with higher- or lower-than-average exposures.

There exists no gold standard data against which to assess the representativeness of exposure levels in IMIS. Therefore, potential differences between exposure levels in IMIS and the broader U.S. population have mostly been investigated through studying associations between IMIS exposure levels and ancillary information describing the circumstances associated with a measurement. Recent examples include the study of Henn *et al.* (2011), which showed that smaller sized and unionized establishments had a

higher probability of having sample results that exceeded the PEL for lead. Assuming this association holds true in the broader universe of occupational exposures, and to the extent that smaller, unionized establishments are over- or under-represented in the IMIS database, the measurements of lead in the IMIS database may not accurately represent the broader set of occupational exposures without adjustment for these variables.

A literature review by Lavoue *et al.* (2013) concluded that no associations of appreciable magnitude were consistently reported across studies between exposure levels in IMIS and variables associated with workplace selection. However, these studies focused on different variables, specific subsets of data, various chemical agents, and different analysis techniques, complicating their interpretation. The main objective of our study was to conduct the most thorough possible evaluation of the association between exposure levels in IMIS with ancillary information across a broad range of chemicals.

3.3 Methods

The Integrated Management Information System (IMIS)

The IMIS databank contains historical information about inspections and associated occupational health and safety enforcement activities conducted by both federal OSHA and state plan inspectors throughout the U.S. The quantitative measurements taken by OSHA to determine compliance with PELs are a specific subset of the databank. Along with measured values, each record contains information about the company, including the name and address of employer, total number of workers at the worksite, and whether it is unionized. The industry is identified by a SIC code from the 1987 or 1972 Standard Industrial Classification (OSHA, 2014c) and also by a code from the North American Industry Classification System (NAICS) after 1997 (OSHA, 2014a). Information on the monitored job is provided in free text form. The date, type of inspection (i.e. complaint, referral, follow-up, planned), scope of inspection (full or partial), type of sample (i.e. area, bulk, personal, screening), and whether the inspection is conducted by federal OSHA or under an OSHA state plan are also recorded. State plans must set workplace safety and health standards that are "at least as effective as" federal OSHA standards (OSHA, 2015b), although they have the option to promulgate additional or more stringent regulations covering hazards not addressed by federal OSHA.

Data preparation

An IMIS extract containing air monitoring results of all compliance evaluations involving sampling from 1979 to 2011 was obtained through the Freedom of Information Act. The extract contained 851,987 records corresponding to 107,647 inspections, covering 1,050 SIC codes and 1,054 agents. Records corresponding to area (n = 35,617), bulk (n = 24,789), blood (n = 128), urine (n = 34), wipe (n = 10,344), screening samples (n = 158), and exact duplicate samples (n = 1,950) were excluded. We also excluded measurements of noise (n = 192,935), as well as measurements that were not analyzed (n = 3,029) or recorded as invalid samples (n = 3,964). The analysis was restricted to all chemical agents that had at least 500 samples. Agent codes without a Threshold Limit Value (TLV) (ACGIH, 2014) or PEL value (OSHA, 2014b), discontinued before 1990 or introduced after 2000 were further excluded (9 agents corresponding to a total of 16,046 records).

Ancillary information examined

We selected eight of the 17 variables recorded in the IMIS exposure dataset (type and scope of inspection, union status, federal/state OSHA plan, year of inspection, SIC code, OSHA region, and type of exposure), plus two variables derived from the publicly available IMIS violation dataset (US Department of Labor, 2014b) (size of establishment, and total penalties assessed as a result of violations at each establishment) (Table I). Variables were excluded if they were either recorded in a non-standardized form, had all entries set to 'missing', or had majority of data attributed to one category (>95%) (Table I footnote d). Prior to modelling, correlations between independent

variables were evaluated using Cramer's V and contingency coefficient (Fisher and van Belle, 1993).

Type of inspection is a record of the reason for conducting each inspection. Only five of the twelve possible categories – planned, complaint, referral by a safety compliance officer or other source, follow-up, and monitoring – were retained for analysis since they represented more than 95% of the data. The 'follow-up' and 'monitoring' categories were combined since both correspond to inspections where the compliance officer returned in a particular facility. The number of workers in each inspected facility was used as a measure of the size of the establishment. This variable was treated as categorical by breaking down the data in tertiles based on the number of workers in each establishment in the whole IMIS exposure dataset (1 to 35 workers = small, 36 to 150 = medium, more than 150 = large). Treating this variable as continuous was also explored in preliminary analyses and did not change the results. The type of exposure variable was divided into two categories: time-weighted average (TWA), corresponding to 8-hr shift-long measurements, and 'short-term', corresponding to short-term exposure level (STEL), peak, or ceiling levels (no difference between these three short-term categories was observed in preliminary analyses).

Coding industry

Industry was included in our analysis in order to control for potential confounding in estimating the effects of other variables related to the representativeness of IMIS

exposure data. A partial aggregation of data across the 4-digit SIC categories was performed so that approximately 30 industry groups were obtained for each chemical agent in the analysis. This yielded a manageable number of categories and still allowed variation across a significant number of groups. For each agent, when fewer measurement results than a predetermined cut point was available for a 4-digit SIC category, the more specific digit was dropped to create a broader category. Cut points were determined by dividing the total number of measurements for the specific agent by 60. The cut points varied from 15 for the least frequent agent to 949 for the most frequent agent. The process was repeated until the number of measurements in the category was greater or equal to the cut point, or the code was reduced to a SIC major division (1-digit). Finally, if a 1-digit code was not associated with more measurements than the cut point, it was put into an 'other' category. This system of classifying SIC codes is typical of those employed when working with other similar analyses (Lavoue *et al.*, 2008; Lee *et al.*, 2015).

Total amount of penalty

The total penalty associated with each establishment was obtained by linking the IMIS exposure dataset with the IMIS violation dataset (linking procedure, based on inspection number and character matching of establishment names, is available from the corresponding author). For a particular measurement, this total was adjusted by excluding fines that came from citations we judged the most closely related to compliance with PEL (Occupational Safety and Health Standards from 1910.1000 through 1910.1052) for the

corresponding inspection. OSHA applied a penalty reduction structure for size allowing for penalty reduction between 10 and 40% for employers with 250 employees or less until 2012 (OSHA has since increased the maximum employer size reduction factor from 40% to 60%) (OSHA, 2015a). The penalty variable was analysed as a four-level categorical variable ('no penalty' category plus 3 categories based on tertiles of the non-zero values). Treating the penalty as a continuous variable or further standardizing by the actual number of employees did not affect the results.

Statistical modeling

Treatment of non-detectable (ND) measurements in IMIS has been discussed by several authors (Melville and Lippmann, 2001; Henneberger *et al.*, 2004; Henn *et al.*, 2011; Lavoue *et al.*, 2013). It has been suggested that an unknown proportion of ND results may correspond to 'not present' situations, i.e. agent was absent from the workplace, with the remaining representing situations where exposure was 'present but not detected' by the sampling and analytical approach. Not knowing which interpretation is closer to the truth, we analysed IMIS data with two different approaches which represent the two extremes of treatment of ND results: all as 'present but not detected' or all as 'not present'. Consistent findings for a predictor across these two approaches would provide coherent evidence for the role of the predictors independently of the methodological issues associated with each approach.

We first modeled the probability of exposure above the TLV (or PEL) following approach of Hamm and Burstyn (2011) and Lee *et al.* (2015). This analysis corresponds to the 'present but not detected' interpretation of NDs and has the added benefit of allowing the inclusion of all data with ND results classified as below the threshold; we refer to this as 'logistic model'. Secondly, we analysed only the detected concentrations using 'linear model'. This analysis assumes that all NDs reflect 'not present' situation, and, as Froines *et al.* (1990) mentioned, should be removed since zero exposure would not be a valid measure in workplaces where the agent is present. Moreover reports have shown that multiple agents are sometimes measured on the same sample media (Henn *et al.*, 2011; Lavoue *et al.*, 2013). This situation would create ND results each time the inspector is interested in a particular agent in the group creating results for the other agents clearly "not present". While the analysis of probability of exceeding exposure threshold may seem less prone to bias, the analysis of detected results with linear models has the advantage of estimating association with levels of exposure measurable on a continuous scale.

Logistic models

For each chemical agent, the measurement results were dichotomized into a binary variable indicating whether or not the measurements exceeded the TLV or OSHA PEL for agents without a TLV. Binomial generalized additive models (GAM; Zuur *et al.* (2009)) were used to evaluate association between all variables and a sample result exceeding the TLV (or PEL). GAM models helped capture non-linear exposure time

trends for benzene in a recent study (Friesen *et al.*, 2012). We applied the current TLV (or PEL) to all years, so that direct comparisons could be made across all years using the same metric. Curvilinear relationship of year with probability of exposure above threshold were modeled using thin-plate splines (Zuur *et al.*, 2009) to ensure for control of confounding in the estimation of relationships between exposure levels and other variables.

Linear models

As discussed in Lavoue *et al.* (2013) the status of a measurement coded as a ND is provided in the same variable that identifies a sample as a shift-long or short-term measurement. This implies that it is not possible to know whether a ND was measured as TWA, STEL, peak, or ceiling value, which prevented the inclusion of this variable in the logistic analyses. We used generalized additive mixed models (GAMM; Zuur *et al.* (2009)) to assess the association of all variables, including the type of exposure (TWA/short-term), on log-transformed detected concentrations for each chemical agent. Agents with >50% of non-detects were further excluded for this analysis (N = 21) since we considered that modeling datasets where more than half the data are excluded likely would not be informative. Before modeling, measurements reported as below one tenth of the empirical limit of detection (calculated as the median value of the first two percent of the distribution) or above the immediately dangerous for life and health (IDLH) value (NIOSH, 2014) were excluded for all chemical agents (max of 1.8% and median of 1.1% of data excluded across agents). These exclusions were carried out since the aim of the

study was to pick out general tendencies, and not identify extreme situations. Finally, short-term measurement values were also excluded for metals and dusts as majority of detected results corresponded to TWA (i.e. 93%). The visit by an OSHA inspector was fitted as a random effect to account for within-inspection correlation based on previous evidence (Lavoue *et al.*, 2008; Henn *et al.*, 2011). Curvilinear relationship of year with concentrations was estimated using separate non-parametric smooth terms for TWA and short-term samples with the same smooth class as in the binomial GAM models.

Model averaging

For both modeling approaches, the main-effect association of each explanatory variable with the response variable was evaluated using model averaging (Burnham and Anderson, 2002; Lavoue *et al.*, 2008; Sauve *et al.*, 2012, 2013). Inference in this approach is based on a set of candidate models (the model set) instead of a single 'final' model obtained by adding and removing variables. This procedure therefore does not assume that a single model is useful, and takes into account uncertainty in model selection. For each chemical agent, the model set was built by first creating sub-models containing all possible combinations of the predictor variables found in Table I. Year and industry were included in all the sub-models to ensure control of confounding. This resulted in a preliminary list of 128 unique model structures for each chemical agent. Coefficients (effect sizes) and standard errors for each predictor variable were calculated by averaging the estimated coefficients from sub-models in which each term appears and

weighting values according to the models' Akaike information criterion (AIC) (Burnham and Anderson, 2002).

For each chemical agent, the association of each predictor with exposure was quantified as follows: for the logistic models, odds ratios (ORs) were calculated by computing the exponential of the multimodel averaged regression coefficients. For a particular variable the OR for a category represents the odds of being higher than TLV for that category over the odds being higher than TLV for the reference category. For the linear models, relative indices of exposure (RIE; Lavoue *et al.* (2006)) were calculated by computing the exponential of the multimodel averaged coefficients.

$$RIE_{levelA}(\%) = 100 * \exp(Coeff_{levelA} - Coeff_{levelRef})$$

where RIE_{levelA} is the relative index of exposure for level A of the variable in question, $Coeff_{levelA}$ is the estimated coefficient corresponding to the category A, and $Coeff_{levelRef}$ is the estimated coefficient corresponding to the reference category. $Coeff_{levelRef}$ is 0 when the reference category is included in the intercept. As an illustration, a 30% RIE for a category means that, on average, exposures in this category are approximately one-third those in the reference category.

Meta-analytic summary measures

For each predictor variable, ORs and RIEs are obtained for every chemical agent. Meta-analytic methods were used to combine results from all chemical agents and get an overall picture of the effect of each predictor (van Houwelingen *et al.*, 2002; Borenstein

et al., 2010). Random effects models were applied to calculate the meta-analytic summary estimates and associated 95% confidence intervals (CI) and forest plots were used to visualize the results. Random-effects modeling approach was selected since it allows the conclusions to be generalized to a wider array of situations than fixed-effect modeling. Heterogeneity of effects was addressed by presenting their magnitudes on forest plots rather than relying on the under-powered test of heterogeneity (Borenstein *et al.*, 2010).

Software

All analyses were performed using the R 3.0.3 statistical software (R Development Core Team, Vienna, Austria), with the packages *mgcv* (Wood, 2014) for binomial GAM modeling, *gamm4* (Wood and Scheipl, 2014) for GAMM modeling, and *metafor* (Viechtbauer, 2014) for the meta-analysis. Model averaging was performed using R functions developed within our research group and are available from the corresponding author.

3.4 Results

Descriptive analysis

The statistical analyses included 511,047 measurements from 66,827 inspections for the period 1979 to 2011. There was a median of 14,549 measurements entered each year in IMIS (interquartile range = [11,673; 19,851]), with a peak of 29,979 samples in 1990 (Figure 1).

Seventy-seven agents were selected for analysis, constituting 91% of all personal samples (23 metals and their compounds, 22 organic solvents, 11 gases, 9 dusts/fibers, 4 isocyanates, and 8 other agents). The most frequently measured agents were lead (n = 56,920), iron oxide fume (n = 30,959), and crystalline silica (n = 26,253) (Table II). The proportion of sample results below the LOD ranged from 3% (wood dust, softwood/hardwood - carbon dioxide) to 91% (antimony).

Logistic models

Regarding multicollinearity, all independent variable pairs had weak correlation (median $r<0.4$ across agents), except for the inspection type/inspection scope pair which had moderate correlation (median r=0.58; IQR=[0.52;0.63] across agents). A correlation of 0.7 is a usual threshold used to flag potential multicollinearity. Across the 77 chemical agents, the median proportion of the variability explained by the full model (i.e. including all tested variables) was 24% (interquartile range = [16%; 38%]).

Table III shows the observed pooled influence of all predictor variables on the odds of exposure exceeding TLV, stratified by group of agents. To illustrate how agent-specific associations relate to the pooled estimate, we present forest plot for solvents and metals (Figure 2A-B). It shows the typical picture of more extreme point estimates being associated with worse precision, as well as highlights general consistency of the direction of effect across agents. Forest plots for each level of each predictor variable are available as Supplementary Appendix 1 at *Annals of Occupational Hygiene* online.

Six variables had meta-analytic ORs away from the null for at least one category: inspection type, OSHA plan, OSHA region, amount of penalty, union status, and establishment size. This means that a high level of consistency across chemical agents was observed for the effect of these variables. For type of inspection, follow-up visits had the highest odds of having a sample result exceed the TLV and this relation was observed across majority of agents (chemical-specific OR in same direction as meta-analytic OR for 61/77 agents). Complaint inspections were associated with higher odds of a sample result exceeding the TLV, but only for metals. Measurements collected under state OSHA plans were less likely to have a sample result exceed the TLV compared to measurements collected by federal OSHA, but this relation was mainly seen for solvents and dusts. An increase in the total amount of penalty was associated with higher odds of having a sample result exceed the TLV and this relation was seen for a majority of agents (62/77 for high penalty compared to none). Substantial differences were seen between

OSHA regions as some regions had odds of a sample result exceeding the TLV close to two times higher than others.

Finally, visual assessment of smoothed curves for each individual agent suggested a decreasing time-trend in the probability of a sample result exceeding the TLV for 50 of the 77 agents (smoothed curves are available as Supplementary Appendix 2 at *Annals of Occupational Hygiene* online).

Linear models

Across the 56 chemical agents, the median proportion of the variability of log-transformed detected concentrations explained by the fixed effects of the full model (i.e. including all tested variables) was 20% (interquartile range = [16%; 24%]). Adding a random-effect structure to the models with visit number as the random effect improved the model fits (Akaike criterion) for all agents. Agent-specific intra-class correlation between measurements taken during the same visit varied between 0.38 and 0.82 (median = 0.66) across the 56 agents.

Table IV shows the observed association of all fixed effects variables with exposure levels, stratified by group of agents. As presented for logistic models, agent-specific RIEs related to the pooled estimate are presented on forest plot for solvents and metals (Figure 2C-D).

Exposure type, OSHA plan, and amount of penalty had the strongest pooled association with levels of detected exposure. For exposure type, exposure levels for shift long TWA samples were lower than short term samples for all chemical agents. For OSHA plan, measurements collected under state OSHA plans had lower levels of detected exposure than measurements collected by federal OSHA for a majority of solvents (chemical-specific RIE in same direction as meta-analytic RIE for 19/21) and all dusts (8/8). An increase in the total amount of penalty was associated with an increase in exposure levels for a majority of agents (47/56 for high penalty compared to none). For type of inspection, exposure levels measured during follow-up inspections were higher than planned inspections for a majority of metals (10/14) and dusts (7/8). Finally, visual assessment of smoothed curves for each individual agent suggested a decreasing time-trend for 53 of the 56 agents for shift long exposure measurements, and for 24 of the 29 agents for short term exposure measurements (smoothed curves are available as Supplementary Appendix 2 at *Annals of Occupational Hygiene* online).

3.5 Discussion

To our knowledge, this study represents the first comprehensive effort to examine the association between reported levels in IMIS and all information available on the circumstances associated with an inspection. We used measurements from 77 chemical agents, covering 1979-2011 and representing 91% of the IMIS chemical exposure dataset. The meta-analysis allowed creating an overall picture of the effect of all predictors (Tables III and IV), and the detailed results for the 77 agents are presented in Supplementary Appendices at *Annals of Occupational Hygiene* online.

Given the difficulty in interpreting the non-detected values coded in the IMIS exposure dataset as 'not present' or 'present but not detected', two separate statistical modeling approaches each corresponding to one of the previous interpretations, were implemented. Modeling the probability of exceeding the TLV allowed the inclusion of samples recorded as ND values, while modeling exposure levels of detected concentrations allowed the inclusion of the variable 'exposure type' (TWA/short-term) and to model levels of exposure measurable on a continuous scale. Consistency in findings for the same variable across these approaches would indicate robustness and provide coherent evidence for the role of these specific variables. In our analysis, directions of effect were generally similar regardless of the approach, showing effects independently of the methodological issues associated with each approach.

The proportion of variance explained by the complete set of variables for the logistic and linear models was relatively small with respective medians of 24% and 20%, but

comparable to other studies on multi-industry data (Coble *et al.*, 2001; Melville and Lippmann, 2001; Lavoue *et al.*, 2008; Lavoue *et al.*, 2011).

The linear models included a random effect structure. As seen in previous analyses of measurements in IMIS for specific chemical agents (Teschke *et al.*, 1999; Lavoue *et al.*, 2008; Henn *et al.*, 2011), a considerable within-visit correlation in all 77 chemical agents' datasets was observed. It was not possible to include such structure in the logistic models since it resulted in problems of complete separation (Albert and Anderson, 1984) of data points for many agents.

Type of inspection was related to exposure in logistic models for all groups of agents and linear models for metals and dusts, with generally higher exposure measured during follow-up inspections than planned inspections. This result reflects the selection bias where inspectors are more likely to return to the poorer performing sites. If the change would have been assessed for follow up inspections paired to the corresponding initial inspection, we would have expected a reduction in exposure levels, although exposures would still be on average higher than in the general working population. Although this link can be established in theory, we were not able to implement it with our dataset. Exposure measured during referral inspections was also generally higher for all groups of agents but only in the logistic models. Our results also suggest the absence of a consistent upward bias linked to visits triggered by complaints by an employee, an aspect which has been suspected since the 1980s and explored many times in previous studies (Froines *et al.*, 1986b; Froines *et al.*, 1990; Gomez, 1997b; Melville and Lippmann, 2001; Lavoue *et*

al., 2008; Henn *et al.*, 2011; Lavoue *et al.*, 2013) without consistent evidence of the presence or absence of such a bias. Our observations point to higher exposures being more likely when an OSHA inspector measures a contaminant in an environment which has already been flagged during a previous inspection visit.

The total amount of penalty associated with an establishment was the most strongly associated with exposure levels in our study. In both the logistic and linear models, our findings were consistent for all groups of agents except for gases and showed that establishments with a history of non-compliance generally have higher exposure levels. Penalties coming from citations related to compliance with PEL (codes from 1910.1000 through 1910.1052) represented a relatively small proportion of total historical penalties. This finding suggests that violative behaviors unrelated to chemical exposure may predict exceedences of the PEL. The relationship between the historical number of citations (some of which do not carry fines) received by an establishment with exposure levels was also tested and yielded similar results. Our observations are however limited by the fact that we did not differentiate between historical penalties issued and those assessed during the inspection in which exposures were measured. Hence 50% of the establishments in the IMIS exposure dataset had experienced only one or two inspections. Furthermore, although we removed citations we judged the most closely related to compliance with PEL for the corresponding inspection, it is possible that citations other than the direct PEL regulations (e.g. risk communication, personal protection, formation, and respirators) are associated with overexposure. Therefore, our variable should be interpreted as a crude proxy or index of total violative behaviors associated with a given

establishment. Despite its limitations, it's the first time it is studied in relation to exposure levels, and we hope it will trigger additional, more refined analyses.

The variable identifying whether the inspection was conducted by federal OSHA or under an OSHA state plan was associated with exposure levels in both the logistic and linear models, with exposures corresponding to federal OSHA being higher than those associated with an OSHA state plan. This finding was very consistent for solvents and dusts, less so for metals. Moreover, substantial differences were also seen between OSHA regions in logistic models. These OSHA region and federal/state differences might correspond to inspection programs that emphasize certain industries with more or less overexposure. However, this would have been mitigated by the inclusion of industry in our analyses, albeit only partly since we did not include interaction between region and industry. It appears more plausible that regional and state/federal compliance and data recording practices are the cause of our observations, such as a possible variation in the recording of non-detected results into the IMIS exposure dataset (Lavoue et al., 2013).

Exposure was marginally lower for large establishments compared to small or medium in both logistic and linear models, with variations across agents where exposure could be increasing (e.g. sulfur dioxide) or decreasing (e.g. asbestos) with establishment size. Other studies have seen an association of establishment size with exposure levels in IMIS (Gomez, 1997b; Melville and Lippmann, 2001; Middendorf, 2004; Henn *et al.*, 2011) also with varying directions of effect.

Limitations

The present study had several strengths, including the analysis of a high proportion of IMIS exposure data, the large sample size for each agent, the significant period of time considered, and application of statistical models that allowed estimation of the simultaneous association of several variables with the concentrations stored in IMIS. Some limitations need to be acknowledged. First, the influence of samples found to be below the limit of detection should be noted. Reports have shown that multiple agents are sometimes measured on the same sample media (Lavoue *et al.*, 2013), precluding users from determining whether a non-detectable result reflected a sample that failed to detect the presence of the agent or a sample where the agent was not the agent being investigated. It is also possible that some inspectors may have only reported concentrations if above a certain action limit near PEL, creating a deficit of true ND in the data. Our strategy of two different analyses (with or without non-detects) permitted to overcome part of this limitation. Availability of a dataset containing all laboratory analyses performed at OSHA's central laboratory (Lavoue *et al.*, 2013), which is the main source of information to the exposure results stored in IMIS, may help understand the mechanism by which records with non-detectable values arise. Second, the lack of information on occupation in a standardized format prevented their inclusion in statistical models, and may have confounded the association between exposure levels and ancillary information. Recent reports created standardized occupation code from the crude job titles provided in IMIS (Burstyn *et al.*, 2014; Russ *et al.*, 2014) and allowed better grouping of the data. This was not possible to implement on the whole databank. Third,

the analysis of variables internal to IMIS, while informative, cannot directly address whether the IMIS data represents exposure levels in the general working population. Although a recent analysis by Lavoue *et al.* (2011) compared exposure levels between IMIS and the French databank COLCHIC for formaldehyde, more external validation efforts may prove helpful.

The statistical models included in this analysis provided quantitative estimates of the effect of 10 variables internal to IMIS. The combined multimodel/meta-analysis method used in this study is a stringent approach which requires a clear signal in order to show a non-null meta-effect. This is clearly illustrated when examining the agent-specific values compared to the meta-coefficients (appendix 1). Conceptually, associations - or lack thereof - seen across a vast majority of agents point towards underlying trends. On the other hand, associations varying across chemicals are somewhat harder to interpret and might result from specific practices for individual agents or be related to other factors which could not be identified or controlled for in this study.

Conclusions

Our analyses revealed consistent associations between exposure levels in IMIS and ancillary variables reflecting the characteristics of establishments selected for inspection and OSHA sampling practices. Higher exposure levels were measured during follow-up inspections than planned inspections. Higher exposure levels were also seen for measurements collected by federal OSHA compared to measurements collected under

OSHA state plans. The exposure levels were also correlated to the amount of historical penalties assessed to a company. These associations are on average of modest size considering the magnitude of random variability that is known to exist in exposures day-to-day and between workers, and are compatibe with other published work (Hall *et al.*, 2002; Friesen *et al.*, 2006; Peters *et al.*, 2011).

Our study suggests that inspection selection processes and targeting practices influence the exposure levels recorded by OSHA. Although the trends we observed cannot be used to directly measure or quantify the extent to which IMIS represents the broader universe of exposures, they should be considered when using IMIS data for any exposure assessment purpose. IMIS is an important source of general information about trends and associations in occupational exposures, given the size and scope of the database and the paucity of other sources of exposure data. Despite its notable limitations, we remain optimistic about the utility of IMIS, and the OIS database that has recently replaced it, in informing occupational exposure assessment in epidemiology and risk assessment.

Acknowledgements

The authors 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. We also thank Robin Ackerman for her insightful comments and suggestions. P.S. was supported by the Institut de recherche Robert-Sauvé en santé et

en sécurité du travail (IRSST). The authors declare that there are no conflicts of interest relating to the material in relation to this paper.

3.6 Tables and figures

Table I : Variables tested in the empirical statistical models

Variable	Description	Type	Number of samples (%)
		Fixed effects	
Inspection type	Reason for conducting inspection	Nominal (4 categories)	
		(1) Planned	159,360 (30)
		(2) Complaint	265,620 (50)
		(3) Follow-up	29,812 (6)
		(4) Referral	72,301 (14)
Inspection scope	Comprehensive or partial survey of the establishment	Nominal (2 categories)	
		(1) Comprehensive	222,606 (42)
		(2) Partial	304,487 (58)
Establishment size	Number of employees working in the establishment	Nominal (3 categories)	
		(1) Small (1-35 employees)	174,108 (33)
		(2) Medium (36-150 employees)	183,260 (35)
		(3) Large (151+ employees)	169,725 (32)
Union status	Union is present or not in the company monitored	Dichotomous	
		(1) Yes	193,198 (37)
		(2) No	333,895 (63)
OSHA plan	Inspection conducted by federal OSHA or under an OSHA state plan	Dichotomous	
		(1) Federal	382,096 (72)
		(2) State	144,997 (28)
OSHA region[a]	Identifies the OSHA region where the inspection took place	Nominal (10 categories)	
		(1) 01_boston	36,890 (7)
		(2) 02_new_york	43,340 (8)
		(3) 03_philadelphia	38,120 (7)
		(4) 04_atlanta	85,138 (16)
		(5) 05_chicago	178,696 (34)
		(6) 06_dallas	50,292 (10)
		(7) 07_kansas_city	26,964 (5)
		(8) 08_denver	30,936 (6)
		(9) 09_san_francisco	12,797 (2)
		(10)10_seattle	23,920 (5)
Penalty	Sum of historical penalties assessed in the establishment monitored	Nominal (4 categories)	
		(1) None	66,322 (13)
		(2) Low	153,452 (29)
		(3) Medium	153,899 (29)
		(4) High	153,420 (29)
Year	Year of sampling	Continuous (integer) 1979 to 2011	
SIC code	Constructed from the aggregation of the 4-digit SIC category	Nominal (median of 30 categories; IQR=[26;33])[b]	
Exposure type[c]	Type of measurement	Dichotomous	
		(1) TWA	300,064 (87)
		(2) Short term	46,644 (13)

Random effect[c]		
Visit number	Identification number for the visit by inspector	Nominal (median of 828 categories; IQR=[297;2172])[b]

[a]https://www.osha.gov/html/RAmap.html
[b]Median number of categories for the variable across chemicals and interquartile range
[c]For linear models of detected levels only
[d]Variables excluded from analysis: advance notice given (majority of data attributed to one category (>95%)), form type (recorded in a non-standardized form), office id (recorded in a non-standardized form), occupation code (all entries set to 'missing'), presence of employee representative (all entries set to 'missing'), interview of employees (all entries set to 'missing'), number of employees exposed (recorded in a non-standardized form), job title (recorded in a non-standardized form), frequency of exposure (recorded in a non-standardized form)

Table II : Descriptive statistics of chemicals in IMIS

Chemical agent	TLV[a] (mg/m^3)	Number of samples	TWA[d] (%)	ST[e] (%)	ND[f] (%)	Fraction of measurements >TLV (%)[g]
Organic solvents						
Toluene	75.4	21,552	53	32	15	24
Xylene	434	13,950	66	11	23	1
2-Butanone	590	6,764	64	10	26	4
Acetone	1190	5,914	68	12	20	2
Methylene Chloride	174	4,876	50	35	15	30
Petroleum naphtha	2000[b]	4,631	63	4	33	0
Isopropyl Alcohol	492	3,916	68	11	21	5
N-Butyl Acetate	713	3,812	67	11	22	0
Stoddard Solvent	525	3,442	70	1	29	4
Hexone	82	3,252	62	11	27	9
Ethyl benzene	87	3,188	62	10	28	3
Benzene	1.6	2,944	28	12	60	13
Methyl Chloroform	1900	2,716	69	15	16	4
Tetrachloroethylene	170	2,631	52	36	12	34
Hexane	177	1,754	74	3	23	10
2-Butoxyethanol	97	1,501	61	1	38	1
Trichloroethylene	54	1,441	51	41	8	65
Phenol	19	1,043	58	1	41	0
N-Butyl alcohol	61	717	46	15	39	7
Ethyl alcohol	1900	698	63	4	33	1
Methyl alcohol	260	692	57	13	30	10
Ethyl acetate	1400	628	77	1	22	0
Metals and their compounds						
Lead, inorganic	0.05	56,920	54	1	45	24
Iron oxide fume	5	30,959	91	1	8	10
Copper fume	0.2	24,742	76	1	23	4
Zinc oxide fume	2	23,964	69	11	20	4
Manganese fume	0.02	23,605	41	40	19	54
Chromium	0.5	19,370	52	1	47	2
Nickel	0.2	17,606	38	1	61	4
Cobalt	0.02	14,329	23	0	77	6
Beryllium	0.00005	13,633	8	7	85	11
Antimony	0.5	12,964	9	0	91	1
Molybdenum	3	11,753	16	0	84	0
Vanadium Fume	0.05	11,467	5	8	87	2
Cadmium fume	0.01	9,962	20	8	72	7
Chromic acid	0.1[b]	3,762	39	17	44	10
Cadmium dust	0.01	3,193	45	4	51	17
Arsenic	0.01	2,746	59	0	41	16
Lead, Inorg.	0.05	1,643	66	0	34	9
fume&dust	1	1,541	83	1	16	11

Copper dusts	0.01	1,014	66	1	33	35
Silver	0.025	881	72	10	18	52
Mercury	15[b]	586	76	0	24	0
Aluminum	1	583	78	1	21	31
Aluminum oxide	2	500	36	0	64	1
Tin						
Gases						
Carbon monoxide	28.8	11,204	78	14	8	37
Formaldehyde	0.37	8,870	49	25	26	21
Hydrogen chloride	3	1,420	12	37	51	12
Ammonia	18	1,164	59	16	25	9
Sulfur dioxide	0.66	908	63	13	24	56
Nitrogen dioxide	0.38	862	10	36	54	34
Ethylene oxide	1.8	829	49	18	33	16
Hydrogen fluoride	0.41	715	39	9	52	21
Carbon dioxide	9000	640	90	7	3	34
Vinyl chloride	2.56	634	25	7	68	10
Chlorine	1.5	503	17	35	48	14
Dust / fibers						
Silica, quartz	0.025	26,253	76	1	23	74
PNOR (total dust)	10	18,264	94	1	5	18
PNOR (resp dust)	3	11,521	93	0	7	15
Asbestos (all forms)	0.1[c]	10,374	39	13	48	26
Silica, cristobalite	0.025	2,203	15	1	84	10
Wood dust, hardwood	1	1,183	95	2	3	71
Fluoride	2.5	894	64	0	36	4
Carbon black	3.5	773	82	1	17	10
Wood dust, softwood	1	557	93	4	3	64
Isocyanates						
4,4'-MDI	0.051	4,221	6	33	61	18
2,4'-TDI	0.036	1,943	11	28	61	13
HDI	0.034	1,498	26	18	56	13
2,6'-TDI	0.036	723	16	8	76	7
Other						
Styrene	85	8,979	53	38	9	60
Welding fumes	5	3,882	93	1	6	18
Oil mist	5	2,831	76	1	23	1
Coal tar pitch vol	0.2	2,531	73	1	26	29
Sulfuric acid	0.2	1,812	49	1	50	12
Sodium hydroxide	2	1,361	48	13	39	1
Nitric acid	5	869	37	6	57	2
Phosphoric acid	1	845	24	2	74	2

[a]2014 ACGIH threshold limit value (TLV).
[b]2014 OSHA permissible exposure limit (PEL).
[c]fibers/cc.
[d]Percentage of values reported as time-weighted average (TWA) data.
[e]Percentage of values reported as short-term data.
[f]Percentage of values reported as nondetects (ND).
[g]Percentage of values above the TLV; the statistic includes both short-term and TWA data because of limitations in the databank design.

Table III : Summary meta-analytic odds ratios (ORs) of a sample result exceeding the TLV for selected variables, stratified by group of agents

Variable/ category	Overall N[a] = 77 OR (95% CI)	Solvents N = 22 OR (95% CI)	Metals N = 23 OR (95% CI)	Gases N = 11 OR (95% CI)	Dusts N = 9 OR (95% CI)
Inspection type					
Planned	1.00 (reference)[b]	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)
Complaint	1.05 (1.00;1.10)[c]	1.04 (0.98;1.11)	**1.15 (1.05;1.26)**	1.01 (0.90;1.13)	0.93 (0.83;1.05)
Follow-up	**1.61 (1.44;1.81)**	**1.65 (1.13;2.40)**	**1.64 (1.36;1.97)**	**1.26 (1.04;1.53)**	**1.80 (1.39;2.32)**
Referral	**1.16 (1.09;1.24)**	1.19 (1.00;1.41)	**1.24 (1.10;1.40)**	1.17 (1.00;1.36)	1.12 (1.00;1.25)
OSHA plan					
Federal	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)
State	**0.82 (0.73;0.92)**	**0.56 (0.44;0.73)**	1.05 (0.90;1.23)	0.84 (0.66;1.08)	**0.77 (0.59;0.99)**
Inspection scope					
Comprehensive	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)
Partial	0.97 (0.94;1.00)	0.96 (0.89;1.04)	1.01 (0.97;1.05)	0.99 (0.91;1.08)	**0.86 (0.80;0.91)**
Union status					
No	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)
Yes	**0.93 (0.89;0.98)**	**0.88 (0.79;0.99)**	0.99 (0.93;1.06)	0.98 (0.87;1.12)	**0.86 (0.77;0.96)**
OSHA region[d]					
01_boston	**1.19 (1.08;1.31)**	1.07 (0.95;1.19)	**1.25 (1.02;1.53)**	**1.35 (1.02;1.79)**	1.05 (0.76;1.45)
02_new_york	**1.13 (1.02;1.26)**	1.18 (0.98;1.41)	**1.29 (1.05;1.58)**	0.91 (0.57;1.46)	0.93 (0.75;1.15)
03_philadelphia	**1.40 (1.26;1.56)**	**1.45 (1.18;1.77)**	**1.57 (1.29;1.90)**	1.00 (0.70;1.44)	**1.35 (1.11;1.64)**
04_atlanta	**1.10 (1.03;1.18)**	**1.17 (1.03;1.32)**	1.04 (0.91;1.19)	1.04 (0.76;1.41)	1.05 (0.92;1.19)
05_chicago	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)
06_dallas	1.00 (0.92;1.08)	1.06 (0.92;1.21)	1.07 (0.93;1.23)	0.79 (0.56;1.11)	**0.76 (0.63;0.92)**

07_kansas_city	1.08 (0.98;1.19)	**1.26 (1.05;1.51)**	0.94 (0.79;1.12)	1.12 (0.80;1.58)	0.90 (0.60;1.33)
08_denver	**1.14 (1.02;1.27)**	1.03 (0.81;1.31)	1.09 (0.93;1.28)	0.84 (0.59;1.18)	1.23 (1.00;1.50)
09_san_francisco	**1.39 (1.22;1.60)**	1.14 (0.97;1.34)	**1.37 (1.03;1.83)**	1.46 (0.86;2.46)	**1.29 (1.09;1.53)**
10_seattle	**1.69 (1.48;1.93)**	**1.76 (1.29;2.39)**	**1.81 (1.53;2.14)**	1.25 (0.74;2.12)	**1.44 (1.14;1.82)**
Establishment size					
Small (1-35)	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)
Medium (36-150)	1.01 (0.97;1.05)	0.98 (0.90;1.06)	1.06 (1.00;1.13)	0.95 (0.82;1.10)	0.96 (0.89;1.05)
Large (151+)	**0.92 (0.87;0.98)**	0.87 (0.75;1.00)	0.99 (0.90;1.10)	0.90 (0.68;1.20)	0.86 (0.71;1.03)
Penalty					
None	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)
Low	**1.27 (1.19;1.35)**	**1.24 (1.10;1.40)**	**1.25 (1.12;1.39)**	1.13 (0.88;1.44)	**1.45 (1.30;1.61)**
Medium	**1.46 (1.34;1.60)**	**1.57 (1.28;1.94)**	**1.35 (1.17;1.55)**	1.24 (0.96;1.59)	**1.89 (1.60;2.23)**
High	**1.54 (1.40;1.71)**	**1.65 (1.32;2.08)**	**1.44 (1.23;1.69)**	**1.36 (1.03;1.79)**	**1.94 (1.54;2.43)**

[a]Number of chemical agents.
[b]OR of the reference levels taken as 1.
[c]95% confidence interval.
[d]https://www.osha.gov/html/RAmap.html.
Bold font is used to show significant OR coefficients

Table IV : Summary meta-analytic relative indices of exposure (RIEs) for selected variables, stratified by group of agents

Variable/ category	Overall N[a] = 56 RIE (95% CI)	Solvents N = 21 RIE (95% CI)	Metals N = 14 RIE (95% CI)	Gases N = 7 RIE (95% CI)	Dusts N = 8 RIE (95% CI)
Inspection type					
Planned	1.00 (reference)[b]	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)
Complaint	1.00 (1.00;1.00)[c]	1.00 (0.99;1.01)	1.01 (0.99;1.02)	0.99 (0.94;1.04)	0.98 (0.94;1.01)
Follow-up	**1.06 (1.03;1.09)**	1.01 (0.98;1.05)	**1.12 (1.05;1.19)**	1.10 (0.94;1.29)	**1.31 (1.13;1.52)**
Referral	1.00 (1.00;1.01)	1.00 (0.99;1.01)	1.02 (1.00;1.04)	1.05 (0.95;1.16)	1.01 (0.98;1.05)
OSHA plan					
Federal	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)
State	**0.86 (0.81;0.91)**	**0.78 (0.69;0.89)**	0.96 (0.83;1.12)	0.87 (0.72;1.04)	**0.74 (0.57;0.97)**
Inspection scope					
Comprehensive	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)
Partial	1.00 (0.99;1.01)	0.99 (0.97;1.02)	1.00 (0.98;1.01)	1.00 (0.97;1.03)	0.96 (0.91;1.02)
Union status					
No	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)
Yes	**0.97 (0.95;0.99)**	0.96 (0.92;1.00)	1.00 (0.99;1.01)	1.00 (0.96;1.03)	0.91 (0.83;1.01)
OSHA region[d]					
01_boston	1.00 (1.00;1.00)	1.00 (1.00;1.00)	0.94 (0.88;1.00)	1.01 (0.97;1.04)	1.00 (0.98;1.02)
02_new_york	1.00 (1.00;1.00)	1.00 (1.00;1.00)	**0.91 (0.86;0.97)**	0.97 (0.91;1.05)	1.00 (0.99;1.01)
03_philadelphia	1.00 (1.00;1.00)	1.00 (1.00;1.00)	**1.06 (1.01;1.12)**	1.00 (0.97;1.03)	1.03 (0.99;1.06)
04_atlanta	1.00 (1.00;1.00)	1.00 (1.00;1.00)	0.97 (0.93;1.01)	1.00 (0.98;1.02)	1.02 (1.00;1.05)
05_chicago	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)
06_dallas	1.00 (1.00;1.00)	1.00 (1.00;1.00)	**0.94 (0.89;0.99)**	1.00 (0.97;1.02)	1.00 (0.98;1.03)

07_kansas_city	1.00 (1.00;1.01)	1.00 (0.99;1.02)	1.07 (1.00;1.14)	1.10 (0.94;1.29)	1.02 (0.98;1.07)
08_denver	1.00 (1.00;1.00)	1.00 (1.00;1.00)	**0.89 (0.80;0.99)**	1.00 (0.96;1.03)	1.01 (0.97;1.05)
09_san_francisco	1.00 (1.00;1.00)	1.00 (1.00;1.00)	0.94 (0.88;1.00)	1.00 (0.95;1.04)	1.01 (0.97;1.05)
10_seattle	1.00 (1.00;1.00)	1.00 (1.00;1.00)	1.05 (0.99;1.12)	0.86 (0.71;1.05)	1.01 (0.95;1.08)
Establishment size					
Small (1-35)	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)
Medium (36-150)	1.00 (0.99;1.00)	1.00 (0.99;1.01)	1.00 (0.99;1.01)	1.00 (0.98;1.02)	0.99 (0.97;1.01)
Large (151+)	**0.98 (0.96;0.99)**	0.97 (0.93;1.00)	0.99 (0.96;1.01)	0.99 (0.93;1.05)	0.92 (0.84;1.02)
Penalty					
None	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)
Low	**1.06 (1.04;1.09)**	1.01 (0.99;1.02)	**1.07 (1.01;1.13)**	1.08 (0.95;1.24)	**1.22 (1.01;1.47)**
Medium	**1.14 (1.10;1.18)**	**1.07 (1.01;1.13)**	**1.13 (1.02;1.26)**	1.09 (0.93;1.28)	**1.39 (1.02;1.90)**
High	**1.18 (1.13;1.23)**	**1.09 (1.02;1.16)**	**1.16 (1.04;1.30)**	1.08 (0.95;1.24)	**1.47 (1.02;2.11)**
Exposure type					
Short term	1.00 (reference)	1.00 (reference)	-	1.00 (reference)	-
TWA	**0.37 (0.34;0.42)**	**0.38 (0.35;0.42)**	-	**0.35 (0.23;0.52)**	-

[a]Number of chemical agents.
[b]RIE of the reference levels taken as 1.
[c]95% confidence interval.
[d]https://www.osha.gov/html/RAmap.html.
Bold font is used to show significant RIE coefficients

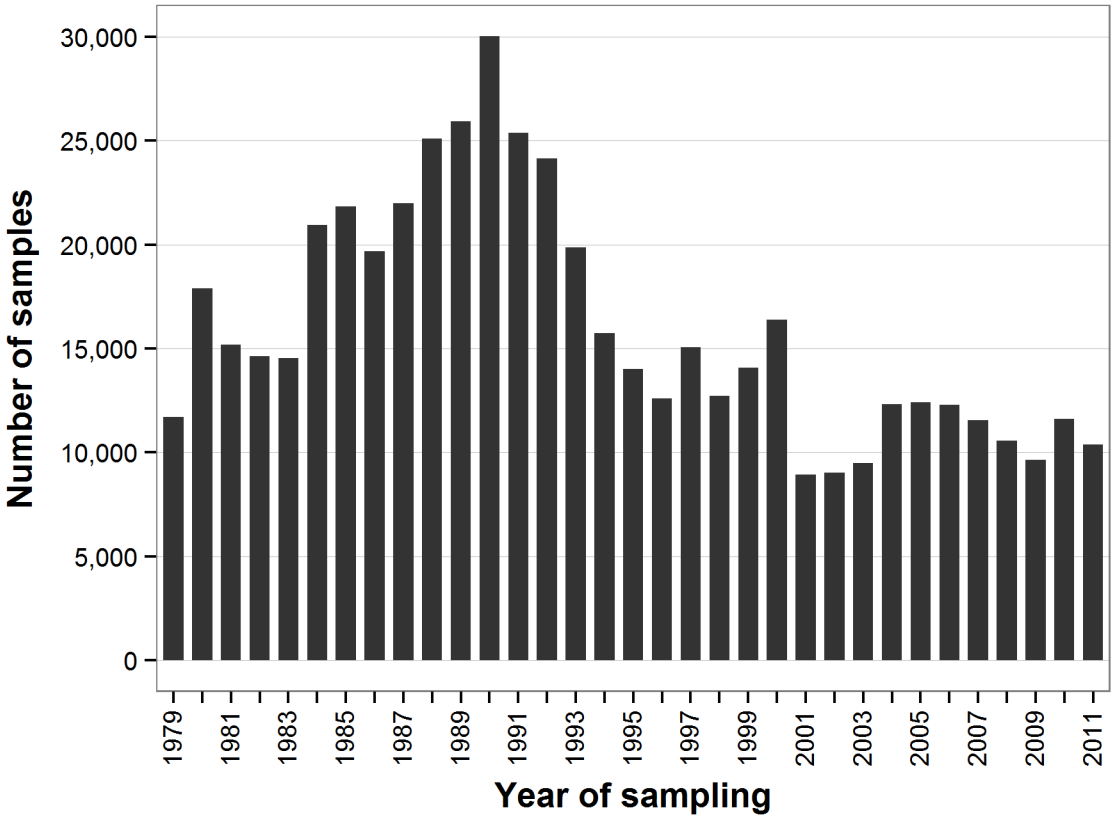


Figure 1 : Number of samples per year in IMIS

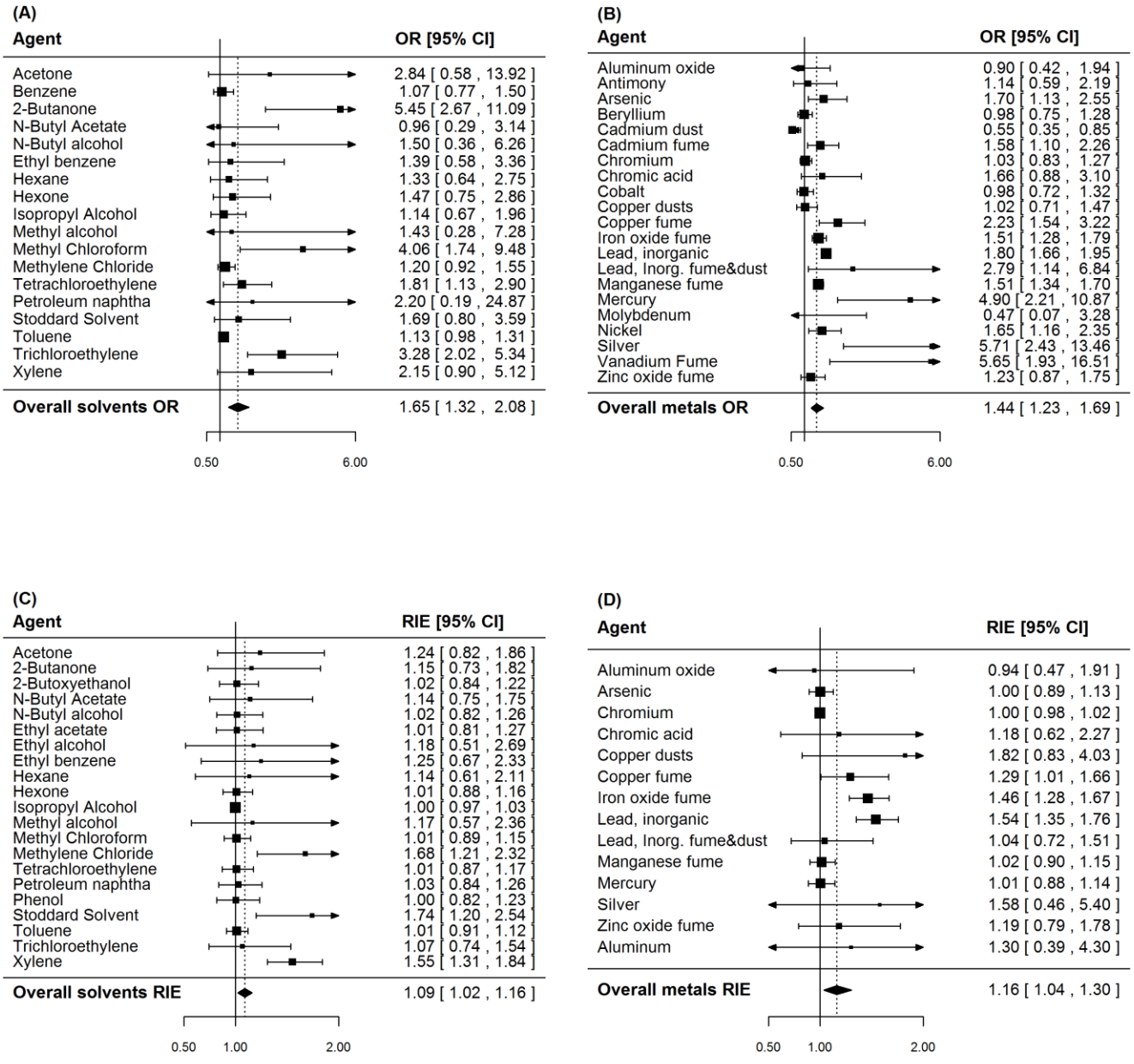


Figure 2 : Agent-specific and meta-analytic ORs and RIEs for 'high' penalty compared to 'none' (A- solvents ORs, B- metals ORs, C- solvents RIEs, D- metals RIEs). Agent-specific ORs and RIEs were pooled with the random effects method. Squares represent agent-specific risk estimates (size of the square reflects the agent-specific statistical weight); horizontal lines, the 95% CIs; diamond, the summary risk estimate and its corresponding 95% CI

3.7 References

ACGIH. (2014) 2014 TLVs and BEIs. ACGIH. 978-1-607260-72-1

Albert A, Anderson JA. (1984) On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. Biometrika; 71: 1-10.

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

Burnham KP, Anderson DR. (2002) Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. New York, NY: Springer. ISBN-13: 978-1441929730.

Burstyn I, Slutsky A, Lee DG *et al.* (2014) Beyond crosswalks: reliability of exposure assessment following automated coding of free-text job descriptions for occupational epidemiology. Ann Occup Hyg; 58: 482-92.

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.

Fisher L, van Belle G. (1993) Biostatistics: a methodology for the health sciences. New York: 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.

Friesen MC, Demers PA, Spinelli JJ *et al.* (2006) From expert-based to quantitative retrospective exposure assessment at a Soderberg aluminum smelter. Ann Occup Hyg; 50: 359-70.

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. (1986a) Occupational health surveillance: a means to identify work-related risks. Am J Public Health; 76: 1089-96.

Froines JR, Wegman DH, Dellenbaugh CA. (1986b) An approach to the characterization of silica exposure in U.S. industry. Am J Ind Med; 10: 345-61.

Gomez MR. (1993) A proposal to develop a national occupational exposure databank. Appl Occup Environ Hyg; 8: 768-74.

Gomez MR. (1997a) Commentary: Recommendations for optimizing the usefulness of existing exposure databases for public health applications. Am Ind Hyg Assoc J; 58: 181-2.

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

Hall AH, Teschke K, Davies H *et al.* (2002) Exposure levels and determinants of softwood dust exposures in BC lumber mills, 1981-1997. AIHA J (Fairfax, Va); 63: 709-14.

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.

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.

Koh DH, Bhatti P, Coble JB *et al.* (2012) Calibrating a population-based job-exposure matrix using inspection measurements to estimate historical occupational exposure to lead for a population-based cohort in Shanghai, China. J Expo Sci Environ Epidemiol; 24: 9-16.

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.

Lavoue J, Gerin 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. (2006) Statistical modelling of formaldehyde occupational exposure levels in French industries, 1986-2003. Ann Occup Hyg; 50: 305-21.

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.

Lee D, Lavoue J, Spinelli J *et al.* (2015) Statistical modeling of occupational exposure to polycyclic aromatic hydrocarbons using OSHA data. J Occup Environ Hyg; 1-14.

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.

Melville R, Lippmann M. (2001) Influence of data elements in OSHA air sampling database on occupational exposure levels. Appl Occup Environ Hyg; 16: 884–99.

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

NIOSH. (2014) Chemical Listing and Documentation of Revised IDLH Values http://www.cdc.gov/niosh/idlh/intridl4.html (Accessed 2 november 2014).

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.

Olsson AC, Gustavsson P, Kromhout H *et al.* (2011) Exposure to diesel motor exhaust and lung cancer risk in a pooled analysis from case-control studies in Europe and Canada. Am J Respir Crit Care Med; 183: 941-8.

OSHA. (2014a) NAICS Manual - The North American Industry Classification System, a 6-digit industry grouping system developed in cooperation with Canada and Mexico. https://www.osha.gov/oshstats/index.html (Accessed 2 november 2014).

OSHA. (2014b) Permissible Exposure Limits – Annotated Tables. https://www.osha.gov/dsg/annotated-pels/ (Accessed 2 november 2014).

OSHA. (2014c) SIC Manual - detailed information for a specified SIC, Division, or Major Group. https://www.osha.gov/oshstats/index.html (Accessed 2 November 2014).

OSHA. (2015a) Annual Review and Scheduled Modification to OSHA's Interim Administrative Penalty Policy. https://www.osha.gov/dep/enforcement/admin_penalty_mar2012.html (Accessed 2 August 2015).

OSHA. (2015b) State Plans - Office of State Programs. https://www.osha.gov/dcsp/osp/ (Accessed 2 August 2015).

Peters S, Vermeulen R, Portengen L *et al.* (2011) Modelling of occupational respirable crystalline silica exposure for quantitative exposure assessment in community-based case-control studies. J Environ Monit; 13: 3262-8.

Rajan B, Alesbury R, Carton B *et al.* (1997) European proposal for core information for the storage and exchange of workplace exposure measurements on chemical agents. Appl Occup Environ Hyg; 12: 31-9.

Russ DE, Ho KY, Johnson CA *et al.* (2014) Computer-Based Coding of Occupation Codes for Epidemiological Analyses. Proc IEEE Int Symp Comput Based Med Syst; 2014: 347-50.

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.

Sauve JF, Beaudry C, Begin D *et al.* (2012) Statistical modeling of crystalline silica exposure by trade in the construction industry using a database compiled from the literature. J Environ Monit; 14: 2512-20.

Sauve JF, Beaudry C, Begin D *et al.* (2013) Silica exposure during construction activities: statistical modeling of task-based measurements from the literature. Ann Occup Hyg; 57: 432-43.

Stewart PA, Rice C. (1990) A source of exposure data for occupational epidemiology studies. Appl Occup Environ Hyg; 5: 359-63.

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.

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

US Department of Labor. (2014b) OSHA Enforcement Data. http://ogesdw.dol.gov/views/data_summary.php (Accessed 15 october 2014).

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

Viechtbauer W. (2014) Meta-Analysis Package for R. Available at http://cran.r-project.org/web/packages/metafor/index.html (Accessed 15 november 2014).

Wood S. (2014) Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation. Available at http://cran.r-project.org/web/packages/mgcv/index.html (Accessed 15 september 2014).

Wood S, Scheipl F. (2014) gamm4: Generalized additive mixed models using mgcv and lme4. Available at http://cran.r-project.org/web/packages/gamm4/index.html (Accessed 15 september 2014).

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.

Zuur A, Ieno EN, Walker N *et al.* (2009) Mixed Effects Models and Extensions in Ecology with R. Springer Science & Business Media. ISBN-13: 978-0387874579.

CHAPITRE 4- Characterization of the selective recording of sample results in OSHA's IMIS databank, 1984-2009: statistical modeling of ancillary information across 78 chemicals

Characterization of the selective recording of sample results in OSHA's IMIS databank, 1984-2009: statistical modeling of ancillary information across 78 chemicals

Philippe Sarazin (1,5), Igor Burstyn (2), Laurel Kincl (3), Jérôme Lavoué (4,5)

(1) Institut de recherche Robert-Sauvé en santé et en sécurité du travail, Chemical and Biological Hazards Prevention, Montréal, Québec, Canada

(2) Drexel University, Environmental and Occupational Health, Philadelphia, Pennsylvania, United States

(3) Oregon State University, College of Public Health and Human Sciences, Corvallis, Oregon, United States

(4) University of Montreal Hospital Research Centre, Montréal, Québec, Canada

(5) Université de Montréal, Department of Occupational and Environmental Health, School of public health, Montréal, Québec, Canada

Corresponding Author:
Philippe Sarazin
Institut de recherche Robert-Sauvé en santé et en sécurité du travail
505, Boul. de Maisonneuve Ouest
Montréal, QC, Canada
H3A 3C2

Email: philippe.sarazin@irsst.qc.ca
Phone: (514) 288-1551 (Ext. 402)

Philippe Sarazin a contribué de façon majeure à la conception et au design de l'étude, a été responsable de l'analyse des données, de l'interprétation et de la présentation des résultats, de la rédaction du manuscrit et de l'édition de l'article.

4.1 Abstract

Objectives: The Integrated Management Information System (IMIS) is the largest multi-industry source of exposure measurements available in North America. In 2010, the Occupational Safety and Health Administration (OSHA) released a second databank, the Chemical Exposure Health Data (CEHD), which contains analytical results of samples collected by OSHA inspectors. However, the two databanks only partially overlap. We investigated the selective recording of sample results into IMIS from CEHD.

Methods: This analysis was based on personal exposure measurements of 78 agents from 1984-2009. The association between 9 variables (level of exposure coded as detected vs. non-detected (ND), panel status, sampling time, issuance of a citation, presence of other detected levels in inspection, year, OSHA region, amount of penalty, and establishment size) and a CEHD sample record being present in IMIS was analyzed using modified Poisson regression.

Results: A total of 588 818 CEHD sample results were examined. The overall proportion of CEHD sample results recorded into IMIS was 38% (50% for detected and 29% for ND measurements). Higher probability of recording of detected vs. ND measurement depended on whether it was part of a panel (risk ratio (RR) = 1.70, 95% confidence interval (CI): 1.67–1.73) or single determination of an agent (RR = 1.24, 95% CI: 1.21–1.26). Probability of recording increased from 1984 to 2009 for measurements of a single agent, but remained constant for samples measured on panels. Some OSHA regions had

probability of recording two times higher than others. None of the other variables were associated with a CEHD sample record being present in IMIS.

Conclusions: Our results indicate that the under-reporting of measurements in IMIS is differential: ND samples (especially the panel ND samples) seem less likely to be recorded in IMIS than other samples. The degree of bias that this selective recording indicates remains to be understood.

4.2 Introduction

Reliable exposure assessment is essential in prevention and management of workplace risks. Availability of information on intensity, duration, and frequency of exposure to chemical air contaminants is critical for a variety of activities including introduction of exposure surveillance programs (Gomez, 1997; Ruttenber *et al.*, 2001; LaMontagne *et al.*, 2002), epidemiologic research (Friesen *et al.*, 2012; Peters *et al.*, 2012; Fritschi *et al.*, 2015; Taeger *et al.*, 2015), and to support the development of predictive models of workplace exposure (Fransman *et al.*, 2011; van Tongeren *et al.*, 2011). In each of these disciplines, readily available sources of measurements that represent a wide variety of chemical agents, industries, and time periods are needed to help estimate exposure accurately.

Multi-industry occupational exposure databanks (OEDBs) have been suggested by several authors as potentially useful sources of individual exposure measurements (Gabriel, 2006; Scarselli *et al.*, 2007; Mater *et al.*, 2016). Set up in several countries in the early 1980s, these databanks contain large quantities of measurement data generated by governmental agencies during various regulatory and prevention activities.

In the U.S., the Occupational Safety and Health Administration (OSHA) maintains two separate databases 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), is the largest multi-industry source of exposure measurements available in North America. This

databank contains air sampling exposure concentrations from surveys performed by OSHA inspectors to verify compliance with Permissible Exposure Limits (PELs). Along with exposure concentrations, it contains detailed information on every inspection visit and company inspected. IMIS has been used by several authors 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 first described in detail by Lavoue *et al.* (2013). It was made available in 2010, and contains the analytical results and associated details of the measurement collected by OSHA inspectors. OSHA officers perform calculations using the sample results (e.g. a time-weighted average concentration, TWA, calculated from several sequential samples on a worker), and record the result of their calculation in IMIS.

Recent studies have reported the combined use of both IMIS and CEHD data to produce a general portrait of exposure levels to polycyclic aromatic hydrocarbons (Lee *et al.*, 2015) and asbestos (Cowan *et al.*, 2015), and to monitor workers' chemical exposure inside plants in the U.S. (Finger and Gamper-Rabindran, 2013). 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. A literature review by Lavoue *et al.* (2013) concluded that no associations of appreciable magnitude were

consistently reported across studies between exposure levels in IMIS and variables associated with workplace selection. A recent report by Sarazin *et al.* (2016) investigated associations between exposure levels and ancillary variables in IMIS across 77 chemical agents, representing >90% of the IMIS dataset. The results also suggested that although certain elements of OSHA's process of selecting worksites for inspection influence the exposure levels that OSHA inspectors encounter, these associations are on average of modest size.

A limitation of most studies having reported on potential bias in IMIS is that analysis was restricted to variables internal to this databank. Although this approach allows determining whether variables related to inspection selection processes and targeting practices are associated with exposure levels, it cannot directly address whether the IMIS data represents exposure levels in the broader US working population. The availability of the CEHD databank provides the opportunity to ask a new question relevant to the potential difference between data in IMIS and data in the general working population: is there a difference between the population of situations sampled by OSHA officers and the population of results recorded in IMIS?

Two OSHA reports published in the 1980s (Mendeloff, 1984; Jones *et al.*, 1986) mentioned that not all measurements made by OSHA officers resulting in lab samples in CEHD ended up causing the creation of a record in IMIS. These findings were recently corroborated in recent studies based on IMIS and CEHD data (Lavoue *et al.*, 2013; Cowan *et al.*, 2015; Lee *et al.*, 2015). Moreover, Lavoue *et al.* (2013) showed that the

proportion of non-detects in the CEHD dataset was higher than in the IMIS dataset for lead, suggesting a differential process: non-detects seem less likely to be recorded in IMIS than other samples. The phenomenon of under-reporting complicates the interpretation of IMIS exposure results as they relate to exposures in the US working population, especially if it appears related to exposure levels.

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. Specifically, we used statistical modeling to investigate the associations between a CEHD sample result being present or not in IMIS and a number of variables reflecting characteristics of that measurement.

4.3 Methods

The OSHA databanks: Integrated Management Information System (IMIS) and Chemical Exposure Health Data (CEHD)

The IMIS exposure databank was accessed through a Freedom of Information Act request. It contains measurement data from all chemicals evaluated during OSHA inspections and information about the company inspected (e.g. name and address of employer, total number of workers at the worksite, whether it is unionized). Date, sample number, sample type (i.e. area, bulk, personal, screening), and type of inspection (i.e. complaint, referral, follow-up, planned) are also recorded.

The CEHD databank was accessed online (OSHA, 2015b). It was made available in 2010, and supplements the IMIS data with the sampling duration, analytical method, and presence of other substances on the same sampling media. The Salt Lake Technical Center, created in 1984, processes all samples collected by the federal and some of the samples collected by State OSHA inspectors.

Data preparation

The IMIS and CEHD extracts available for this analysis contained data from the time periods 1979–2011 and 1984–2009, respectively. The comparison was therefore restricted to measurements from both databanks taken between 1984 and 2009. The IMIS extract contained 851,987 records corresponding to 107,647 inspections, covering 1,054

agents. Cleaning of IMIS data was described in Sarazin *et al.* (2016). Briefly, records corresponding to area, bulk, blood, urine, wipe, screening samples, noise, and exact duplicate samples were excluded. The CEHD online dataset contained 1,908,373 records corresponding to 1,082 agents. Cleaning of CEHD data was described in detail in appendix 1 in Lavoue *et al.* (2013). Briefly, records were removed if they were not personal measurements, irrelevant for exposure assessment (e.g. blank samples), had uninterpretable misspellings, missing information, or judged erroneous. A detailed description of the data cleaning process and a link to an application that recreates the cleaned data from the online raw files is available in Lavoue *et al.* (2013). The analysis was restricted to all CEHD chemical agents that had at least 1,000 samples. Agent codes without a Threshold Limit Value (TLV) (ACGIH, 2014) or PEL value (OSHA, 2014) were further excluded (2 agents corresponding to a total of 3,255 records).

Linkage between IMIS and CEHD

Both datasets were combined using the 'sampling number' variable which was considered as identifying a unique 'evaluation' made by an inspector (e.g. a worker's full shift). Hence several samples in CEHD for one sampling number would correspond to partial shift measurements aggregated by the inspector prior to recording in IMIS through the calculation of a time weighted average. 'Sampling number' is not a perfect variable for linking CEHD to IMIS, since in some cases aggregated sampling time seems unrealistically high (i.e. >600 minutes), and some records in IMIS have the same sampling number. However, preliminary analyses on three chemical agents (lead,

formaldehyde, and toluene) showed that these issues were relatively infrequent (i.e. <5%). Multiple records tied to a single 'sampling number' in CEHD were therefore treated as sequential partial-shift measurements and aggregated to calculate total sampling time and concentration result 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 reported as a non-detect. A report by Lee *et al.* (2015), based on IMIS and CEHD CTPV measurements, noted uncertainty about the method used by inspector to calculate the TWA concentration results reported in IMIS from the CEHD sequential measurements. In some cases, the exposure levels at periods when sampling did not occur were considered the same as during the time of sampling, but in other cases they were also considered to be zero. Having no information to select one approach over the other we considered the aggregated sampling time as representative of the full shift.

Ancillary information examined

The focus of the study was on identifying the determinants of a CEHD sample result being recorded or not into IMIS. Our primary interest was to investigate if sample recording into IMIS was related to the measured exposure level, e.g. are ND records less likely to be recorded in IMIS than other samples? Or are high exposure levels more likely to be recorded? For each chemical agent, the average concentration results were divided into three categories indicating whether or not the measurements were detected and

whether they exceeded the PEL of the agent at the time of measurement: non-detect (ND), detected below PEL (det<PEL), and equal or above PEL (det>=PEL).

Several reports have indicated that multiple agents are often measured on the same sample media (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 non-detectable result reflected a sample where the agent of interest was not detected or a sample where the agent was not investigated but analyzed nevertheless (e.g. this would typically occur in the case of metals). We suspected that ND samples within a panel, only reported by the lab because of analytical protocol, would therefore tend to not be recorded into IMIS. The variable 'field number' included in the CEHD databank identifies samples collected on the same sampling media. CEHD samples were therefore divided into two categories indicating whether or not the sample was belonging to a panel (panel=yes if more than 1 agent on the sampling media).

We also suspected that CEHD samples 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 exposure results in the CEHD inspection and calculating the proportion that were detected. This variable was analysed 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 used as a measure of the size of the establishment. This variable was treated as categorical by breaking down the data in tertiles based on the number of workers in each establishment in the whole CEHD dataset (1 to 35 workers = small, 36 to 150 = medium, more than 150 = large). Treating this variable as continuous was also explored in preliminary analyses and did not change the results.

The Standard Industrial Classification (SIC) code was not included as a predictor because of the correlation with the level of exposure when using finely defined categories (e.g. high exposures for iron and steel foundries).

We used the publicly available IMIS violation dataset (US Department of Labor, 2014b) to create 4 variables associated with the violative behavior of a given establishment (the linking procedure between CEHD and IMIS violation databanks, based on inspection number and character matching of establishment names, is available from the corresponding author). Past reports on small subsets of IMIS data for particular agents have shown that the decision to record a result in IMIS might depend on the issuance of a citation for overexposure (Mendeloff, 1984; Jones *et al.*, 1986). The variable 'PEL citations' represents the number of citations issued during the inspection we judged the most closely related to compliance with PEL (Occupational Safety and Health Standards (OSH) 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). A recent report by Sarazin *et al.* (2016) also discussed the possibility that

citations other than the direct PEL regulations could be associated with overexposure (e.g. risk communication, personal protection, respirators). 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 not included in the two variables described above (they include OSH standards mainly associated with mechanical and electrical safety issues, and physical hazards). Each of the three previous variables was analysed as a three-level categorical variable ('no' category plus 2 categories based on median of the non-zero values). Finally, we created the 'penalty' variable, representing the total amount of fines historically received by an establishment. This variable was analysed as a four-level categorical variable ('no penalty' category plus 3 categories based on tertiles of the non-zero values). We did not correct penalty amounts according to establishment size because OSHA already applies a penalty reduction structure for size, allowing for reduction between 10 and 40% for employers with 250 employees or less until 2012 (OSHA has since increased the maximum employer size reduction factor from 40% to 60%) (OSHA, 2015a). Treating the penalty as a continuous variable or further standardizing by the actual number of employees did not affect the results. Prior to modelling, correlations between independent variables were evaluated using Cramer's V and contingency coefficient (Fisher and van Belle, 1993).

Statistical modeling

Full dataset

The CEHD sample results were transformed into a binary variable indicating whether or not the measurement also appears in the IMIS databank. This binary variable was the response variable.

For the main analysis, we used Poisson regression using a sandwich variance estimator for estimation of risk ratios with binary outcome (Greenland, 2004; Zou, 2004; Hamm and Burstyn, 2011) to model the probability of a CEHD sample result being recorded into IMIS. We used Poisson rather than logistic regression to prevent overestimation of relative risks, likely to occur for prevalence over 25% (Zou, 2004; Hamm and Burstyn, 2011). All variables described above were added in the model, as well as sample year, sampling duration, OSHA region, and chemical agent code (Table I). 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 lab, but probably not of interest to the OSHA officer. Effect coding was used to code for the OSHA region and chemical agent variables to allow comparisons of the probability of a CEHD sample result being recorded into IMIS in a specific category with the overall mean probability of recording. The reference group for all other variables was chosen a priori as 'none' for none/low/high variables and as 'small' for establishment size variable. The association of year and sampling time with probability of recording were modeled using polynomial functions, with a maximal complexity of 8 degrees. The association of each categorical predictor with recording or not into IMIS was quantified using risk ratios (RRs). For a variable, the RR for a specific category represents the probability of a

CEHD sample result being recorded into IMIS for that category divided by the probability of recording for the reference category of that variable.

A supplementary analysis was performed using the same model structure described above except that it incorporated an interaction of year with exposure level and panel status. This was done to evaluate if the temporal trend of probability of recording differed between detected and ND records, measured alone or as part of panel samples (this analysis was not selected as the main analysis because it complicated the presentation of results). To provide a graphical comparison of the predicted recording probabilities from 1984 to 2009, we calculated predicted probabilities for each year for detected and ND records, measured alone or as part of panel samples. Predictions were based on the median sampling duration and on the reference level of each categorical variable (for OSHA region and chemical agent effect coded variables, categories closest to the population average were chosen, i.e. 05_chicago and 2-Butanone, respectively).

Agent by agent

The previous modelling strategy (i.e. 'full dataset') implies different recording probabilities for each agent, but the same influence of other predictors across agents. To evaluate whether the associations were consistent across chemical agents, we also modeled the probability of recording on an agent by agent basis. The analyses were restricted to chemical agents that had at least 2,500 samples (these agents represent 92% of all CEHD data). The Poisson regression models were fitted to each agent dataset with

the same final structure as the 'full dataset' main analysis (interaction of exposure level with panel sample + same complexity of polynomial functions for year and sampling time). For each predictor variable, RRs were obtained for every chemical agent. Meta-analytic methods were used as in Sarazin *et al.* (2016) to combine results from all chemical agents, assess variability across agents, and get an overall picture of the effect of each predictor (van Houwelingen *et al.*, 2002; Borenstein *et al.*, 2010). Briefly, random effects models were applied to calculate the meta-analytic summary estimates and forest plots were used to illustrate how agent-specific associations relate to the pooled estimate. Heterogeneity of effects was addressed by presenting their magnitudes on forest plots (Borenstein *et al.*, 2010).

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.

4.4 Results

Descriptive analysis

The statistical analyses included 1,034,000 CEHD analytical measurements, which were reduced to 588,818 average concentration results corresponding to 36,442 inspection visits for the period 1984 to 2009. Seventy-eight agents were selected for analysis, constituting 95% of all personal samples (32 organic solvents, 22 metals and their compounds, 6 gases, 6 dusts/fibers, 4 isocyanates, and 8 other agents) (Table II). Fourteen out of the 15 most frequently measured agents in CEHD were metals (n > 13,549); other agents that were frequently measured included particles not otherwise regulated (PNOR) - respirable dust (n = 23,129), silica - quartz (n = 12,978), PNOR - total dust (n = 12,036), and toluene (n = 9,030). The proportion of sample results above the PEL ranged from 0% (11 solvents and 3 metals) to 40% (coal tar pitch volatile). The overall proportion of CEHD sample results recorded into IMIS was 38% (50% for detected records and 29% for ND records). Lead, vinyl chloride, 4,4'-MDI, and phenol had the highest proportion of samples recorded into IMIS (65%, 60%, 59%, 59%, respectively), whereas cadmium (twa), tin, and PNOR - respirable dust had less than 10% of their samples recorded.

Statistical modeling

Main analysis: modeling the full dataset

Regarding multicollinearity, all independent variable pairs had weak correlation based on Cramer's V and contingency coefficient ($r<0.4$), except for the panel sample/chemical agent and exposure level/chemical agent pairs which had moderate correlation ($r=0.66$ and $r=0.48$, respectively). A correlation of 0.7 was our threshold to flag multicollinearity. The associations of year and sampling time with probability of a CEHD sample result being recorded into IMIS were best modeled with fourth degree and first degree (i.e. linear function) polynomials, respectively.

Table III shows the observed influence of all categorical predictor variables on the probability of recording. The three variables that had the strongest association were exposure level, panel sample, and OSHA region. For exposure level, detected sample results were associated with a higher probability of recording into IMIS than ND results, and this relation was seen regardless if measured on panels or alone. Furthermore, the probability of recording was similar between detected samples <PEL and >=PEL on both panel and non-panel samples. However, the contrast between ND and detected results was higher in panel compared to non-panel samples. In panel samples, detected results were 60% (detected<PEL) and 71% (detected>=PEL) more likely to be recorded in IMIS compared to ND samples, whereas in non-panel samples, detected results were 33% (detected<PEL) and 24% (detected>=PEL) more likely to be recorded in IMIS compared to ND samples.

RRs varied between chemical agents with probability of recording varying from 6.7 times lower than average for PNOR - respirable dust, to 2 times higher than average for lead.

The most frequently measured solvents and metals, as well as all gases and isocyanates, had a higher than average probability of recording. Probability of recording varied between OSHA regions from 1.4 times lower than average for 02_new_york, to 1.4 times higher than average for 08_denver. Finally, visual assessment of the smoothed curve for the effect of year of sampling suggested that the probability of recording remained fairly constant from 1984 to 2009 (Figure 1).

Supplementary analysis: adding the interaction of year with exposure level and panel status

Figure 2 shows the predicted probability of recording for year stratified by exposure level and panel sample status (based on the model that incorporated an interaction of year with exposure level and panel status). Visual assessment of smoothed curves suggests an increase in the probability of recording from 1984 to 2009 for both ND (~50% increase) and detected samples (~130% increase) when measured alone, with detected samples showing higher probability of recording through the whole time period. The probability of recording remained fairly constant from 1984 to 2009 for detected samples measured on panels, while a small decrease in probability was seen for ND samples (~20% decrease) in more recent years.

Secondary analysis: modeling agent-by-agent datasets

The secondary analysis looked at the associations of predictors with the probability of recording on an agent by agent basis. Twenty-four agents were included in this analysis. The observed pooled influence of all predictor variables are shown in Table V as meta-analytic risk ratios. Similar to the 'full dataset' approach, the agent-by-agent approach confirmed the association of exposure level, panel sample, and OSHA region with probability of recording. Detected>=PEL sample results were associated with a higher probability of recording into IMIS than ND results regardless if measured on panels (meta-RR = 1.66, 95% confidence interval (CI): 1.46–1.88) or alone (meta-RR = 1.53, 95% CI: 1.35–1.73), and this relation was observed across majority of agents (chemical-specific RR>1.00 for 21/24 agents for panel=no and panel=yes).

In addition, forest plots which illustrate how agent-specific associations relate to the pooled estimate are available for each level of each predictor variable (appendix 1). We show as an example in Figure 3 the agent-specific and meta-analytic RRs for 'detected above PEL' records compared to ND records stratified by panel status. Visual assessment of forest plots generally showed homogeneity across agents for all predictors, with a few exceptions of note: issuance of PEL citations during an inspection was associated with higher probability of recording for cadmium (Twa), while an increase in the proportion of detected samples in the inspection was associated with lower probability of recording for this same agent. An increase in the total amount of penalty was associated with higher probability of recording for acetone, styrene, and silica (quartz).

4.5 Discussion

To our knowledge, this study represents the first comprehensive effort to examine empirically mechanisms by which data is reported to IMIS. We used measurement data from 78 chemical agents, covering 1984-2009 and representing 95% of the CEHD chemical exposure dataset. This study 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 and with the use of statistical modeling to study concomitantly several potential explanatory variables. The main modeling approach on the full CEHD dataset allowed estimating the average effect of each predictor across agents, while modeling data agent-by-agent allowed examining if effect of predictors were consistent across agents.

The overall proportion of CEHD samples recorded into IMIS was 38% for the period 1984-2009, with a higher proportion of recording for detected results (50%) compared to ND results (29%). Lavoue *et al.* (2013), Lee *et al.* (2015), and Cowan *et al.* (2015) found similar proportions of CEHD samples having a link to an IMIS record for lead, coal tar pitch volatiles, and asbestos, respectively.

The results from the multivariate regression model showed that the level of exposure of the sample, as well as its panel status, were the most strongly related to the CEHD sample being recorded into IMIS. Higher probability of recording of detected vs. ND measurement was seen regardless if measured on panels or alone, although the contrast between detected and ND results was higher in panel compared to non-panel samples.

Our findings suggest ND samples might be considered by an OSHA officer as not 'worth' reporting in IMIS. Moreover, ND results from chemicals that are most of the time measured as part of panel samples, such as metals and dusts, would be even less interesting to the officer since they would only be analysed because of analytical protocol, and not be of interest for risk analysis. On the other hand, the probability of recording was generally the same for detected samples regardless if measured alone or on a panel, which is consistent with the fact that detected results measured on panels likely reflect a sample where the agent was being investigated by the officer. Our results also suggest that high exposure levels did not seem to cause more reporting into CEHD, and that it is rather the status of being detected which is the main determinant. These results support the hypothesis of Lavoue *et al.* (2013), Jones *et al.* (1986), and Mendeloff (1984) that the IMIS under-reporting is differential: non-detects seem less likely to be recorded in IMIS than other samples.

Analysis of time trends in the proportions of a CEHD sample result being recorded into IMIS showed that the overall probability of recording remained fairly constant from 1984 to 2009. The supplementary analysis (which included an interaction of year with exposure level and panel status) however suggests that the overall flat curve would be a mix of increase in the proportion of recording for samples measured alone (detected and ND), steadiness for detected samples measured as part of panels, and small decrease for ND samples measured as part of panels. While the increase seen for samples measured alone might be explained by an increased technical ease in data entry over the years, the

steadiness/small decrease seen for samples measured on panels is more difficult to interpret based on the information available.

Samples with higher sampling time were more likely to be recorded into IMIS, although the association was moderate. These observations might be explained by the fact that OSHA inspectors are more likely to record a result in IMIS when efforts were made to monitor a whole working shift or a longer task.

There were substantial differences between chemical agents in the probability of recording. The observed differences between agents 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 issues (e.g. very low probability of recording for frequently measured PNOR – respirable and total dusts). It is probable that most PNOR results in the CEHD databank are only present because the officer requested another agent, such as a metal or silica. This hypothesis is supported by the fact that the crude proportion of recording was 35% for total dust samples measured alone compared to 9% when measured on a panel. There were also substantial differences in the probability of recording between OSHA regions. These differences are likely related to region specific practices regarding how data is reported into IMIS.

Probability of recording was marginally associated with the type of citations issued during the inspection (i.e. PEL, respiratory protection and hazard communication, and

other). It was expected that measurements made during inspections for which citations were issued might be more systematically recorded but a strong signal was not observed. Since the single other study (Jones *et al.*, 1986) that looked at the association between recording of a sample in IMIS and issuance of a citation for overexposure was done on a small sample size extract in the early 1980s, it was not possible to make a meaningful comparison of results. The same was also expected for companies with a history of noncompliance but again the probability of recording was similar regardless of the historical amount of penalty assessed to the establishment.

Regardless of the modeling approach, both the full model and agent-by-agent models provided generally similar direction and amplitude of effects of predictors. The homogeneity of results across agents observed in the agent-by-agent approach confirmed the assumption of similar effects of predictors on probability of recording across agents in the main modeling approach.

Limitations

Some limitations need to be acknowledged. First, the lack of information on the unit monitored (e.g. worker, task, and tool) may have confounded the aggregation procedure used to regroup sequential CEHD samples. Sampling number was used as a single identifier of what we called an 'evaluation', but it's possible that CEHD samples linked to one sampling number could correspond to the monitoring of different workers or tasks.

However, this issue would likely be minor since 97% of the aggregated sampling durations were below the standard 480 minutes working shift.

Second, the interpretation of samples found to be below the limit of detection should be noted. It has been suggested (Henn *et al.*, 2011; Lavoue *et al.*, 2013; Lee *et al.*, 2015; Sarazin *et al.*, 2016) that an unknown proportion of ND results may correspond to 'present but not detected' situations, i.e. agent was present in the workplace but at a low level, with the remaining representing 'not present' situations, i.e. agent was absent from the workplace (true zeros). These reports showed that multiple agents are sometimes measured on the same sample media. This situation would create ND results each time the inspector is interested in a particular agent in the group creating results for the other agents clearly 'not present'. It is probable that ND results corresponding to true zeros (mainly the panel ND results) would be less recorded than NDs corresponding to situations where exposure was 'present but not detected', but it was not possible to investigate this issue from the CEHD and IMIS variables included in this study.

Third, although comparison of the IMIS databank with an external source of data (i.e. CEHD) helped to investigate the mechanisms by which results are recorded in IMIS, CEHD data cannot be seen as a 'random sample' of exposure levels found in the US workplaces. CEHD data instead represent the population of facilities receiving OSHA inspection visits where samples were collected. Comparing IMIS with CEHD thus helped investigating whether IMIS data represents exposure levels in the broader US working population, but this approach cannot be regarded as a true external comparison.

Finally, the fact that short term measurements were treated with the TWA PEL limit instead of the short term or ceiling PEL limit when available may have confounded the classification of detected sample results (i.e. det<PEL, det>=PEL). Some short term measurements may have been classified as above instead of below the PEL since short term limits are higher. However, this issue seems minor since only 6 of the 15 agents in the study currently with a short term PEL limit had more than 15% of results with a sampling time below 30 minutes.

Conclusions

IMIS and CEHD monitoring data are important elements for occupational exposure assessment. Recent reports (Lavoue *et al.*, 2013; Sarazin *et al.*, 2016) concluded that inspection selection processes and targeting practices influence the exposure levels recorded by OSHA. Findings in the current study improve the knowledge base on the issue of how representative OSHA data is towards the broader US working population. The under-reporting of ND results 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 sample result in IMIS. Although these trends cannot be used to directly measure the extent to which IMIS represents the broader US population, they should be considered when using IMIS data for any exposure assessment purpose. The conclusions found in

this study are mainly 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 is therefore of paramount importance to determine whether the phenomenon of under-reporting also exists with state datasets.

4.6 Tables and figures

Table I : Variables tested in the empirical statistical models

Variable	Description	Type	Number of samples (%)
Exposure level	Level of exposure of CEHD sample	Nominal (3 categories)	
		(5) Non-detected (ND)	349,859 (59)
		(6) Detected<PEL[a]	212,146 (36)
		(7) Detected>=PEL	26,813 (5)
Panel sample	CEHD sample is part or not of a panel of samples	Nominal (2 categories)	
		(3) No	94,348 (16)
		(4) Yes	494,470 (84)
Year	Year of sampling	Continuous (integer) 1984 to 2009	
Sampling time	Duration of sampling of CEHD sample in minutes	Continuous (integer) IQR=[240;448]	
PEL citations	Number of PEL citations issued during the inspection	Nominal (3 categories)	
		(4) None	279,855 (48)
		(5) Low (1-4 citations)	176,685 (30)
		(6) High (5+ citations)	132,278 (22)
Respiratory protection and hazard communication citations	Number of respiratory protection and hazard communication citations issued during the inspection	Nominal (3 categories)	
		(1) None	219,071 (37)
		(2) Low (1-3 citations)	220,141 (37)
		(3) High (4+ citations)	149,606 (25)
Other citations	Number of other types of citations issued during the inspection	Nominal (3 categories)	
		(1) None	131,976 (22)
		(2) Low (1-5 citations)	260,922 (44)
		(3) High (6+ citations)	195,920 (33)
Detected samples in inspection	Proportion of detected samples in the inspection	Nominal (4 categories)	
		(1) None	31,672 (5)
		(2) Low (1-33%)	231,275 (39)
		(3) Med (34-67%)	258,563 (44)
		(4) High (68-100%)	67,308 (11)
Establishment size	Number of employees working in the establishment	Nominal (3 categories)	
		(1) Small (1-35 employees)	194,925 (33)
		(2) Medium (36-150 employees)	209,657 (36)
		(3) Large (151+ employees)	184,236 (31)
Penalty	Sum of historical penalties assessed in the establishment monitored	Nominal (4 categories)	
		(5) None	49,399 (8)
		(6) Low	179,846 (31)
		(7) Medium	180,043 (31)
		(8) High	179,530 (30)
OSHA region[b]	Identifies the OSHA region where the inspection took place	Nominal (10 categories)	
		(11)01_boston	50,350 (9)
		(12)02_new_york	76,267 (13)
		(13)03_philadelphia	51,146 (9)
		(14)04_atlanta	63,965 (11)

		(15)05_chicago	213,620 (36)
		(16)06_dallas	64,004 (11)
		(17)07_kansas_city	21,306 (4)
		(18)08_denver	32,328 (5)
		(19)09_san_francisco	11,729 (2)
		(20)10_seattle	4,103 (1)
Chemical agent	Identifies the chemical agent sampled by OSHA inspector	Nominal (78 categories)	

[a]Permissible Exporue Limit (PEL) at time of measurement
[b]https://www.osha.gov/html/RAmap.html

Table II : Descriptive statistics of chemicals in CEHD and IMIS

Chemical agent	CEHD				IMIS
	Number of records	ND (%)	Proportion of measurements >=PEL (%)[a]	Proportion of samples recorded into IMIS (%)	Number of records
Organic solvents					
Toluene	9,030	10	1	53	22,063
Xylene	7,987	15	2	52	14,366
2-Butanone	3,355	17	3	49	6,892
Stoddard Solvent	2,734	33	0	54	3,629
Acetone	2,567	15	1	53	6,086
Methylene Chloride	2,492	16	22	53	5,057
Isopropyl Alcohol	2,331	17	2	51	4,037
N-Butyl Acetate	2,163	19	0	52	3,915
Petroleum naphtha	2,087	42	0	53	4,760
Hexone	2,068	23	2	49	3,350
Ethyl benzene	1,725	22	1	53	3,323
Benzene	1,633	65	4	45	3,120
Methyl Chloroform	1,525	12	4	49	2,786
Tetrachloroethylene	1,386	13	9	52	2,709
N-Butyl alcohol	1,319	35	2	26	750
2-Butoxyethanol	1,208	31	0	50	1,560
VM P Naphtha	1,197	39	1	21	549
Ethyl acetate	994	21	1	23	639
Hexane	894	16	0	48	1,783
Ethyl alcohol	826	28	1	34	714
Trichloroethylene	803	12	11	43	1,466
Phenol	697	42	0	59	1,067
Methyl (n-amyl) ketone	540	26	1	18	276
Trimethylbenzene	527	23	6	12	207
Isobutyl Acetate	506	21	0	17	239
Methyl alcohol	480	24	6	47	720
Heptane	439	21	0	23	287
2-Ethoxyethyl Acetate	410	46	0	25	316
Isobutyl Alcohol	406	31	0	21	232
n-Propyl Acetate	308	8	2	25	192
Diacetone Alcohol	300	39	0	25	183
n-Propyl Alcohol	239	20	2	26	192
Metals and their compounds	48,291	58	20	65	59,700
Lead, inorganic	33,663	36	7	42	25,697
Copper fume	33,212	36	1	40	25,023
Zinc oxide fume	33,137	95	0	29	13,612
Antimony	33,035	61	1	36	20,208
Chromium	32,783	96	1	29	14,478

Beryllium	32,641	75	1	33	18,380
Nickel	32,492	15	4	47	32,145
Iron oxide fume	32,410	89	1	31	14,985
Cobalt	32,283	36	0	40	24,716
Manganese fume	32,132	91	1	29	12,070
Vanadium Fume	31,890	90	0	29	12,382
Molybdenum	15,869	88	1	33	10,369
Cadmium fume	13,549	83	5	3	634
Cadmium (Twa)	4,364	52	11	27	3,004
Arsenic	2,834	52	7	48	3,928
Chromic acid	1,622	78	1	9	504
Tin	1,439	49	28	30	1,044
Silver	1,244	45	8	40	1,624
Chromium (VI)	796	18	7	21	1,569
Copper dusts	659	15	17	44	933
Mercury	628	43	7	52	3,334
Cadmium dust					
Gases					
Formaldehyde	4,408	31	6	52	9,173
Hydrogen chloride	869	56	3	55	1,458
Ammonia	863	28	7	48	1,219
Ethylene oxide	635	35	20	45	890
Sulfur dioxide	463	29	5	54	939
Vinyl chloride	358	74	5	60	682
Dust / fibers					
PNOR (resp dust)	23,129	82	2	5	12,092
Silica, quartz	12,978	27	28	53	28,172
PNOR (total dust)	12,036	28	8	23	18,980
Asbestos (all forms)	8,284	62	10	50	11,794
Silica, cristobalite	1,303	97	3	40	2,286
Fluoride	462	56	4	42	951
Isocyanates					
4,4'-MDI	3,058	65	12	59	4,360
2,4'-TDI	1,852	70	2	44	1,967
2,6'-TDI	1,674	65	10	36	736
HDI	1,476	62	10	41	1,557
Other					
Styrene	3,819	9	11	53	9,152
Coal tar pitch vol	1,290	24	40	55	2,591
Sulfuric acid	1,032	50	2	51	1,855
Sodium hydroxide	939	31	2	51	1,401
Nitric acid	572	65	2	52	875
Cyclohexanone	428	24	1	37	469
Naphtha (Coal Tar)	396	50	4	17	339
Methyl Methacrylate	345	35	2	49	467

[a]Percentage of values above the PEL.

Table III : Risk ratios (RR) of a CEHD sample result being recorded into IMIS for all variables

Variable/ category	RR (95% CI)
(Panel_sample) X (Expo_level)	
Panel_no : ND	1.00 (reference)[a]
Panel_no : det<PEL	1.33 (1.31;1.36)[b]
Panel_no : det>=PEL	1.24 (1.21;1.26)
Panel_yes : ND	0.75 (0.74;0.77)
Panel_yes : det<PEL	1.20 (1.18;1.22)
Panel_yes : det>=PEL	1.28 (1.25;1.31)
Sampling time (min)	1.03 (1.03;1.04)[c]
PEL citations in inspection	
None (0)	1.00 (reference)
Low (1-4)	0.97 (0.97;0.98)
High (5+)	0.93 (0.92;0.94)
RespProt & HazComm citations in inspection	
None (0)	1.00 (reference)
Low (1-3)	1.03 (1.02;1.04)
High (4+)	1.05 (1.04;1.05)
Other citations in inspection	
None (0)	1.00 (reference)

Low (1-5)	1.05 (1.04;1.06)
High (6+)	1.02 (1.01;1.03)
Other det_level in inspection	
None (0%)	1.00 (reference)
Low (1-33%)	1.00 (0.99;1.02)
Med (34-66%)	1.03 (1.01;1.04)
High (67%+)	1.01 (0.99;1.03)
Establishment size	
Small (1-35)	1.00 (reference)
Medium (36-150)	1.01 (1.00;1.01)
Large (151+)	0.95 (0.95;0.96)
Penalty	
None	1.00 (reference)
Low	1.00 (0.98;1.01)
Medium	0.97 (0.96;0.99)
High	0.92 (0.91;0.93)
OSHA region[d]	
Mean of OSHA regions	1.00 (reference)
01_boston	1.05 (1.04;1.07)
02_new_york	0.73 (0.72;0.74)
03_philadelphia	0.76 (0.75;0.77)
04_atlanta	0.93 (0.93;0.94)
05_chicago	1.01 (1.00;1.02)
06_dallas	1.13 (1.12;1.14)

07_kansas_city	1.11 (1.10;1.13)
08_denver	1.38 (1.36;1.39)
09_san_francisco	0.84 (0.82;0.86)
10_seattle	1.24 (1.21;1.28)

[a]RR of the reference levels taken as 1.
[b]95% confidence interval.
[c]Corresponds to the RR variation for an increase of one SD (156 minutes).
[d]https://www.osha.gov/html/RAmap.html

Table IV : Risk ratios (RR) of a CEHD sample result being recorded into IMIS for all chemical agents

Chemical agent	RR (95% CI)	Chemical agent	RR (95% CI)
Mean of chemical agents	1.00 (reference)[a]		
Toluene	1.31 (1.28;1.34)	Iron oxide fume	1.23 (1.21;1.25)
Xylene	1.32 (1.28;1.35)	Cobalt	1.14 (1.12;1.16)
2-Butanone	1.17 (1.13;1.22)	Manganese fume	1.14 (1.12;1.16)
Stoddard Solvent	1.42 (1.37;1.47)	Vanadium Fume	1.09 (1.07;1.12)
Acetone	1.33 (1.28;1.38)	Molybdenum	1.10 (1.07;1.12)
Methylene Chloride	1.34 (1.29;1.39)	Cadmium fume	1.20 (1.17;1.23)
Isopropyl Alcohol	1.28 (1.23;1.33)	Cadmium (Twa)	0.12 (0.11;0.13)
N-Butyl Acetate	1.36 (1.31;1.42)	Arsenic	0.83 (0.79;0.87)
Petroleum naphtha	1.47 (1.41;1.53)	Chromic acid	1.35 (1.30;1.41)
Hexone	1.32 (1.26;1.37)	Tin	0.33 (0.28;0.38)
Ethyl benzene	1.39 (1.33;1.46)	Silver	0.87 (0.80;0.94)
Benzene	1.38 (1.30;1.45)	Chromium (VI)	1.13 (1.06;1.22)
Methyl Chloroform	1.16 (1.10;1.22)	Copper dusts	0.56 (0.49;0.64)
Tetrachloroethylene	1.15 (1.09;1.21)	Mercury	1.17 (1.07;1.27)
N-Butyl alcohol	0.72 (0.66;0.79)	Cadmium dust	1.41 (1.31;1.52)
2-Butoxyethanol	1.25 (1.18;1.33)	Formaldehyde	1.30 (1.26;1.34)
VM P Naphtha	0.57 (0.51;0.63)	Hydrogen chloride	1.57 (1.48;1.67)
Ethyl acetate	0.61 (0.54;0.68)	Ammonia	1.19 (1.11;1.28)
Hexane	1.22 (1.14;1.31)	Ethylene oxide	1.19 (1.10;1.30)
Ethyl alcohol	0.91 (0.83;1.00)	Sulfur dioxide	1.35 (1.25;1.46)
Trichloroethylene	1.04 (0.96;1.12)	Vinyl chloride	1.74 (1.60;1.90)
Phenol	1.46 (1.37;1.56)	PNOR (resp dust)	0.15 (0.15;0.16)
Methyl (n-amyl) ketone	0.50 (0.42;0.59)	Silica, quartz	1.36 (1.33;1.38)

Trimethylbenzene	0.32 (0.26;0.40)	PNOR (total dust)	0.59 (0.57;0.61)
Isobutyl Acetate	0.43 (0.35;0.52)	Asbestos (all forms)	1.41 (1.37;1.44)
Methyl alcohol	1.13 (1.03;1.24)	Silica, cristobalite	1.54 (1.44;1.65)
Heptane	0.60 (0.51;0.71)	Fluoride	1.24 (1.12;1.38)
2-Ethoxyethyl Acetate	0.70 (0.59;0.82)	4,4'-MDI	1.73 (1.68;1.79)
Isobutyl Alcohol	0.58 (0.48;0.69)	2,4'-TDI	1.45 (1.38;1.53)
n-Propyl Acetate	0.63 (0.52;0.77)	2,6'-TDI	1.17 (1.10;1.25)
Diacetone Alcohol	0.71 (0.59;0.86)	HDI	1.18 (1.11;1.25)
n-Propyl Alcohol	0.68 (0.55;0.83)	Styrene	1.20 (1.16;1.24)
Lead, inorganic	1.98 (1.96;2.01)	Coal tar pitch vol	1.34 (1.27;1.41)
Copper fume	1.18 (1.16;1.20)	Sulfuric acid	1.41 (1.33;1.50)
Zinc oxide fume	1.12 (1.10;1.14)	Sodium hydroxide	1.30 (1.22;1.39)
Antimony	1.11 (1.09;1.13)	Nitric acid	1.57 (1.45;1.70)
Chromium	1.13 (1.11;1.15)	Cyclohexanone	0.90 (0.80;1.01)
Beryllium	1.12 (1.10;1.15)	Naphtha (Coal Tar)	0.49 (0.40;0.61)
Nickel	1.14 (1.12;1.16)	Methyl Methacrylate	1.22 (1.09;1.35)

[a]Mean absolute probability of a CEHD sample result being recorded into IMIS = 36%

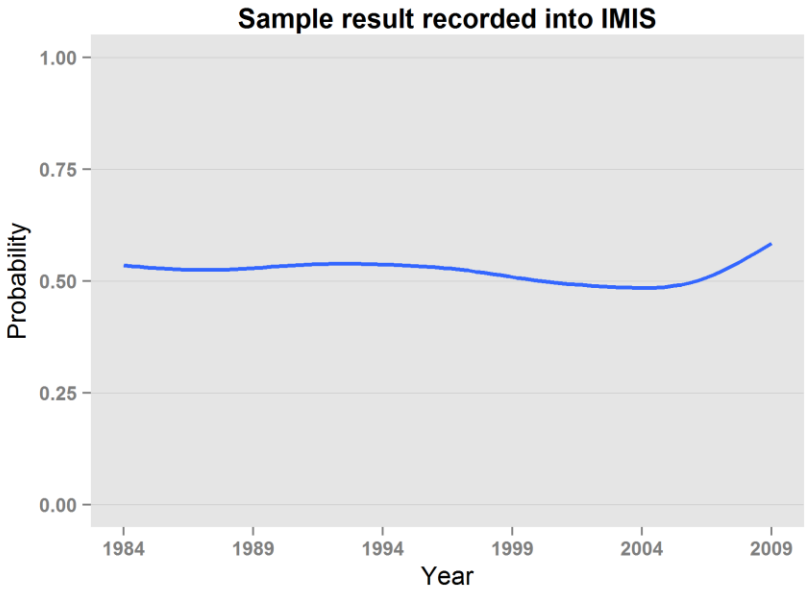


Figure 1 : Risk ratios (RR) of a CEHD sample result being recorded into IMIS for year

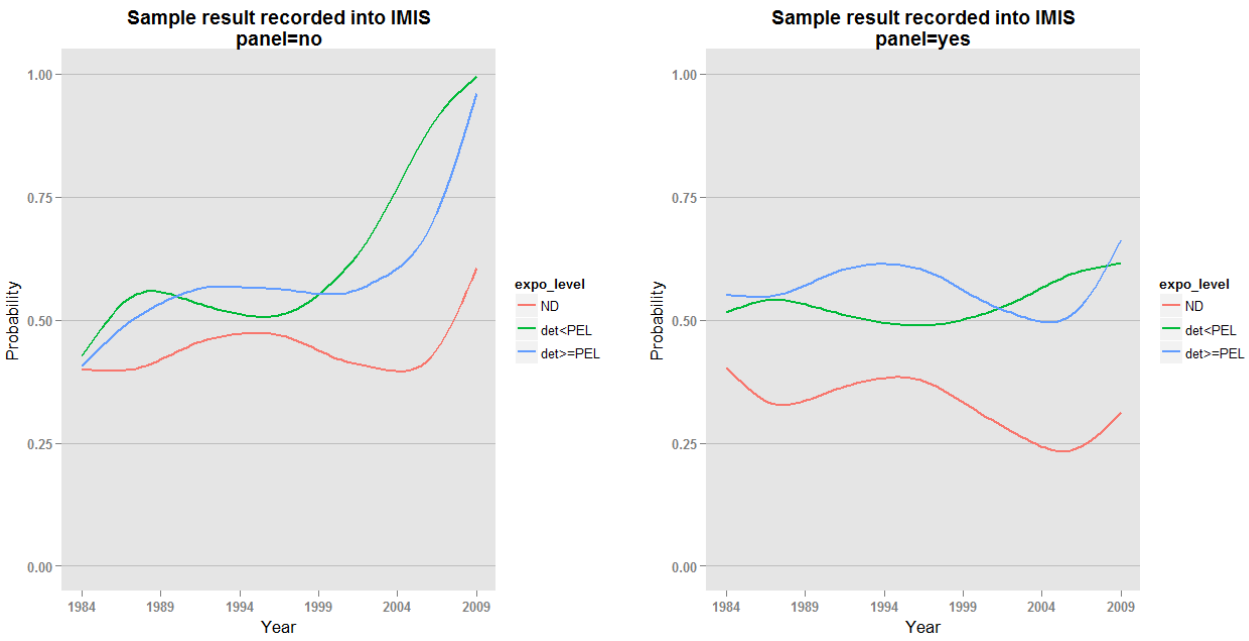


Figure 2 : Risk ratios (RR) of a CEHD sample result being recorded into IMIS for year, stratified by panel status and level of exposure (based on model incorporating an interaction of year with panel status and level of exposure)

Table V : Summary meta-analytic risk ratios (RRs) of a CEHD sample result being recorded into IMIS

Variable/ category	RR (95% CI)
(Panel_sample) X (Expo_level)	
Panel_no : ND	1.00 (reference)[a]
Panel_no : det<PEL	1.37 (1.25;1.49)[b]
Panel_no : det>=PEL	1.53 (1.35;1.73)
Panel_yes : ND	0.82 (0.73;0.92)
Panel_yes : det<PEL	1.22 (1.11;1.35)
Panel_yes : det>=PEL	1.35 (1.19;1.54)
Sampling time (min)	1.04 (1.03;1.05)[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 & HazComm citations in inspection	
None (0)	1.00 (reference)
Low (1-3)	1.02 (1.00;1.04)
High (4+)	1.05 (1.03;1.08)
Other citations in inspection	
None (0)	1.00 (reference)

Low (1-5)	1.04 (1.01;1.06)
High (6+)	1.02 (0.99;1.04)
Other det_level in inspection	
None (0%)	1.00 (reference)
Low (1-33%)	1.02 (0.98;1.06)
Med (34-66%)	0.99 (0.97;1.00)
High (67%+)	0.96 (0.92;1.00)
Establishment size	
Small (1-35)	1.00 (reference)
Medium (36-150)	1.01 (0.99;1.03)
Large (151+)	0.95 (0.94;0.97)
Penalty	
None	1.00 (reference)
Low	1.00 (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)
01_boston	1.22 (1.14;1.32)
02_new_york	0.82 (0.75;0.89)
03_philadelphia	0.88 (0.80;0.98)
04_atlanta	1.07 (0.99;1.16)
05_chicago	1.13 (1.07;1.19)
06_dallas	1.21 (1.16;1.27)

07_kansas_city	1.28 (1.19;1.38)
08_denver	1.48 (1.37;1.60)
09_san_francisco	0.86 (0.82;0.89)
10_seattle	1.15 (1.02;1.31)

[a]RR of the reference levels taken as 1.
[b]95% confidence interval.
[c]Corresponds to the RR variation for an increase of one SD. Median SD of 144 minutes; IQR==[143;165].
[d]https://www.osha.gov/html/RAmap.html

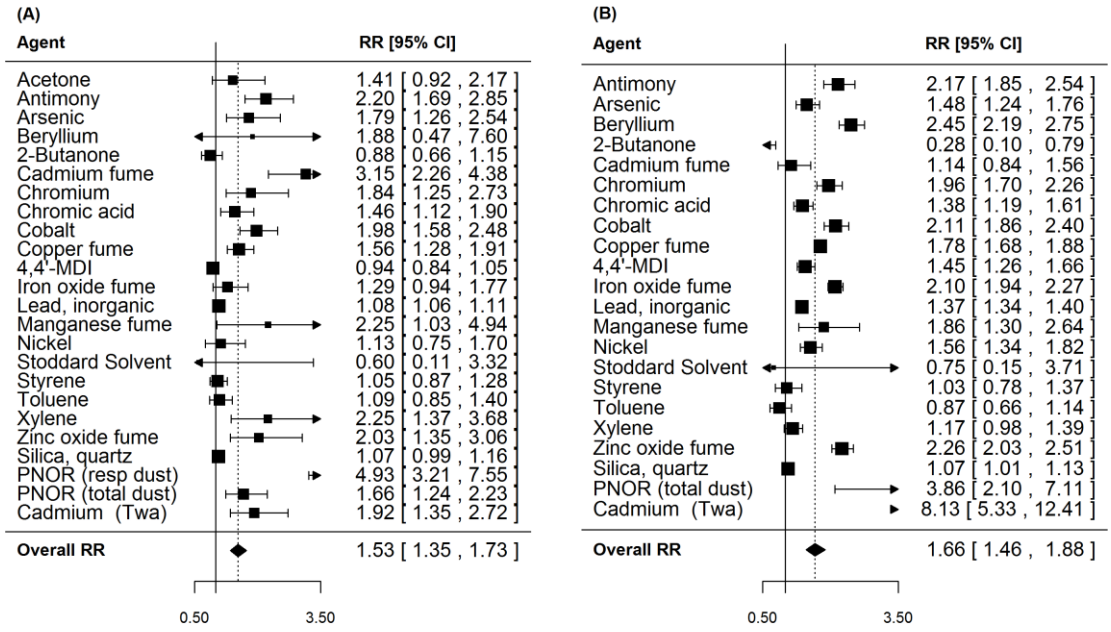


Figure 3 : Agent-specific and meta-analytic RRs of a CEHD sample result being recorded into IMIS for 'exposure level>=PEL' compared to 'exposure level=ND' (A- panel=no, B-panel=yes). Agent-specific RRs were pooled with the random effects method. Squares represent agent-specific risk estimates (size of the square reflects the agent-specific statistical weight); horizontal lines, the 95% CIs; diamond, the summary risk estimate and its corresponding 95% CI

4.7 References

ACGIH. (2014) 2014 TLVs and BEIs. ACGIH. 978-1-607260-72-1

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

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.

Finger SR, Gamper-Rabindran S. (2013) Mandatory disclosure of plant emissions into the environment and worker chemical exposure inside plants. Ecological Economics; 87: 124-36.

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

Fransman W, Van Tongeren M, Cherrie JW *et al.* (2011) Advanced Reach Tool (ART): development of the mechanistic model. Ann Occup Hyg; 55: 957-79.

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.

Gomez MR. (1997) Commentary: Recommendations for optimizing the usefulness of existing exposure databases for public health applications. Am Ind Hyg Assoc J; 58: 181-2.

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 (NTIS No PB2003-104588).

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 D, Lavoue J, Spinelli J *et al.* (2015) Statistical modeling of occupational exposure to polycyclic aromatic hydrocarbons using OSHA data. J Occup Environ Hyg; 1-14.

Mater G, Paris C, Lavoue J. (2016) Descriptive analysis and comparison of two French occupational exposure databases: COLCHIC and SCOLA. Am J Ind Med.

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 november 2014).

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. https://www.osha.gov/dsg/annotated-pels/ (Accessed 2 november 2014).

OSHA. (2015a) Annual Review and Scheduled Modification to OSHA's Interim Administrative Penalty Policy. https://www.osha.gov/dep/enforcement/admin_penalty_mar2012.html (Accessed 2 August 2015).

OSHA. (2015b) Chemical Exposure Health Data. https://www.osha.gov/opengov/healthsamples.html (Accessed 2 August 2015).

Peters S, Kromhout H, Olsson AC *et al.* (2012) Occupational exposure to organic dust increases lung cancer risk in the general population. Thorax; 67: 111-6.

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.

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.

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

US Department of Labor. (2014b) OSHA Enforcement Data. http://ogesdw.dol.gov/views/data_summary.php (Accessed 15 october 2014).

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.

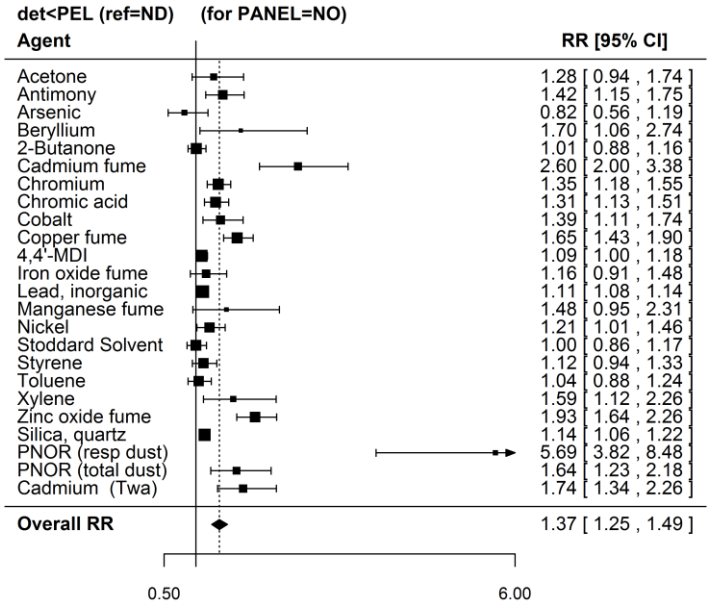
Viechtbauer W. (2014) Meta-Analysis Package for R. Available at http://cran.r-project.org/web/packages/metafor/index.html (Accessed 15 november 2014).

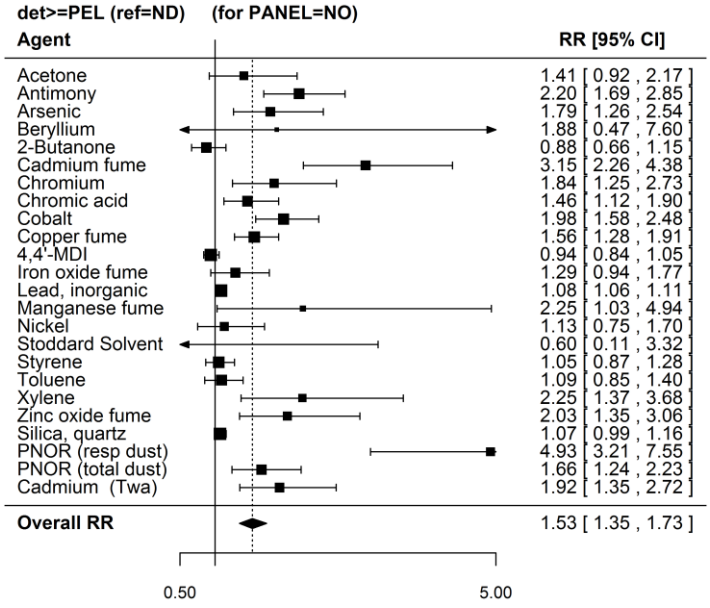
Wickham H. (2015) Package 'ggplot2'. Available at https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf (Accessed 15 november 2014).

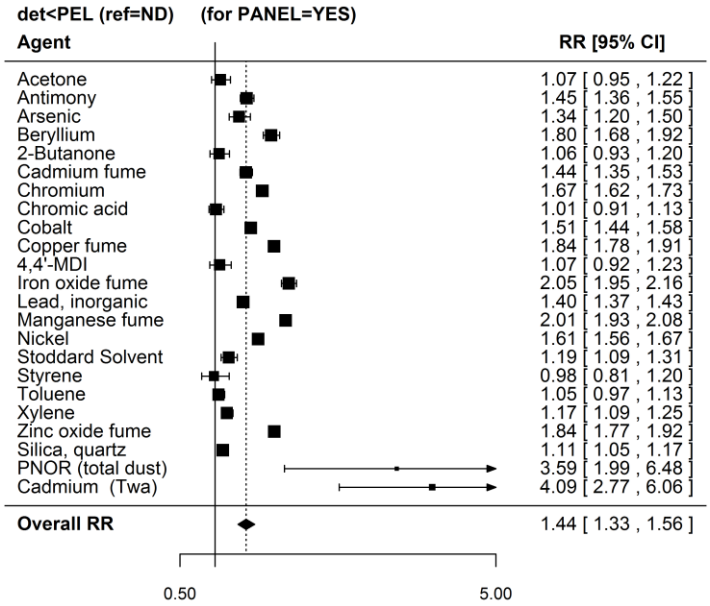
Zeileis A. (2015) Robust Covariance Matrix Estimators. Available at https://cran.r-project.org/web/packages/sandwich/sandwich.pdf (Accessed 15 november 2014).

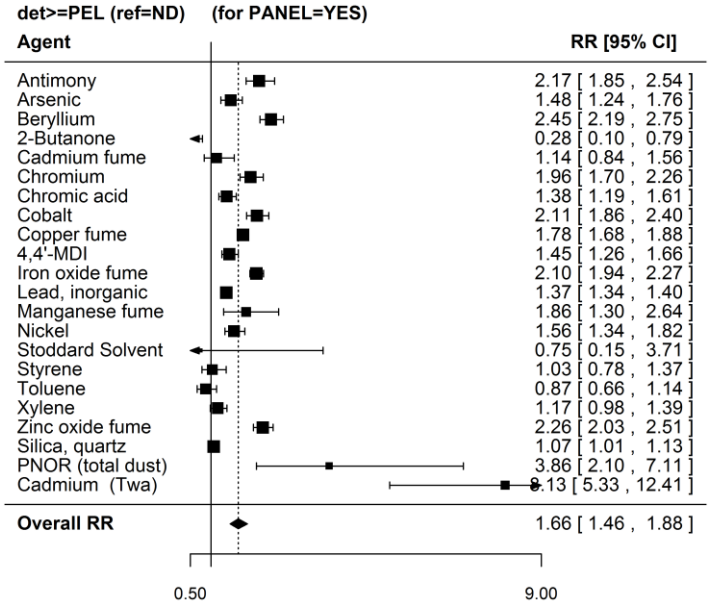
Zou G. (2004) A modified poisson regression approach to prospective studies with binary data. Am J Epidemiol; 159: 702-6.

4.8 Appendix 1

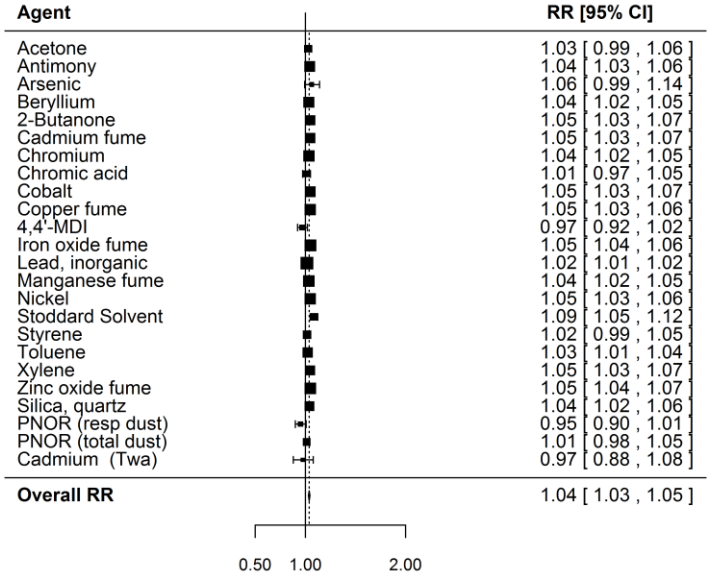




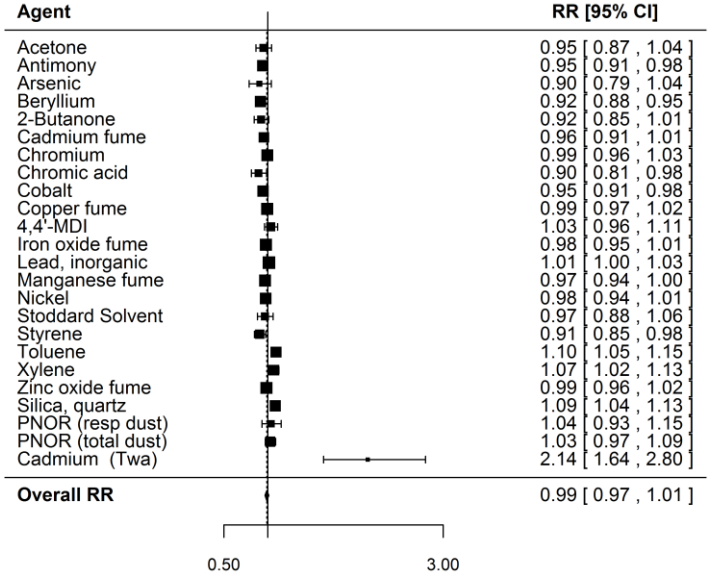

det<PEL (ref=ND) (for PANEL=YES)
Agent
RR [95% CI]
Acetone 1.07 [0.95 , 1.22]
Antimony 1.45 [1.36 , 1.55]
Arsenic 1.34 [1.20 , 1.50]
Beryllium 1.80 [1.68 , 1.92]
2-Butanone 1.06 [0.93 , 1.20]
Cadmium fume 1.44 [1.35 , 1.53]
Chromium 1.67 [1.62 , 1.73]
Chromic acid 1.01 [0.91 , 1.13]
Cobalt 1.51 [1.44 , 1.58]
Copper fume 1.84 [1.78 , 1.91]
4,4'-MDI 1.07 [0.92 , 1.23]
Iron oxide fume 2.05 [1.95 , 2.16]
Lead, inorganic 1.40 [1.37 , 1.43]
Manganese fume 2.01 [1.93 , 2.08]
Nickel 1.61 [1.56 , 1.67]
Stoddard Solvent 1.19 [1.09 , 1.31]
Styrene 0.98 [0.81 , 1.20]
Toluene 1.05 [0.97 , 1.13]
Xylene 1.17 [1.09 , 1.25]
Zinc oxide fume 1.84 [1.77 , 1.92]
Silica, quartz 1.11 [1.05 , 1.17]
PNOR (total dust) 3.59 [1.99 , 6.48]
Cadmium (Twa) 4.09 [2.77 , 6.06]
Overall RR 1.44 [1.33 , 1.56]
0.50
5.00


det>=PEL (ref=ND) (for PANEL=YES)
Agent
RR [95% CI]
Antimony 2.17 [1.85 , 2.54]
Arsenic 1.48 [1.24 , 1.76]
Beryllium 2.45 [2.19 , 2.75]
2-Butanone 0.28 [0.10 , 0.79]
Cadmium fume 1.14 [0.84 , 1.56]
Chromium 1.96 [1.70 , 2.26]
Chromic acid 1.38 [1.19 , 1.61]
Cobalt 2.11 [1.86 , 2.40]
Copper fume 1.78 [1.68 , 1.88]
4,4'-MDI 1.45 [1.26 , 1.66]
Iron oxide fume 2.10 [1.94 , 2.27]
Lead, inorganic 1.37 [1.34 , 1.40]
Manganese fume 1.86 [1.30 , 2.64]
Nickel 1.56 [1.34 , 1.82]
Stoddard Solvent 0.75 [0.15 , 3.71]
Styrene 1.03 [0.78 , 1.37]
Toluene 0.87 [0.66 , 1.14]
Xylene 1.17 [0.98 , 1.39]
Zinc oxide fume 2.26 [2.03 , 2.51]
Silica, quartz 1.07 [1.01 , 1.13]
PNOR (total dust) 3.86 [2.10 , 7.11]
Cadmium (Twa) 8.13 [5.33 , 12.41]
Overall RR 1.66 [1.46 , 1.88]
0.50
9.00

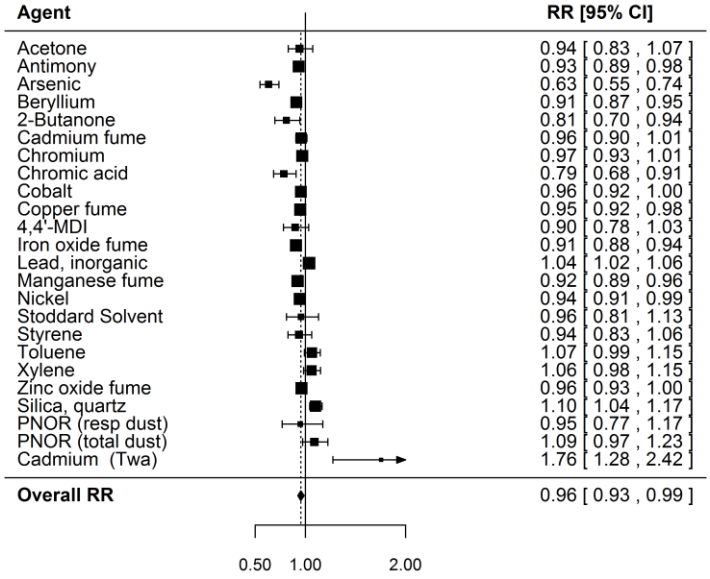
Sampling time (ref=increase of one SD)



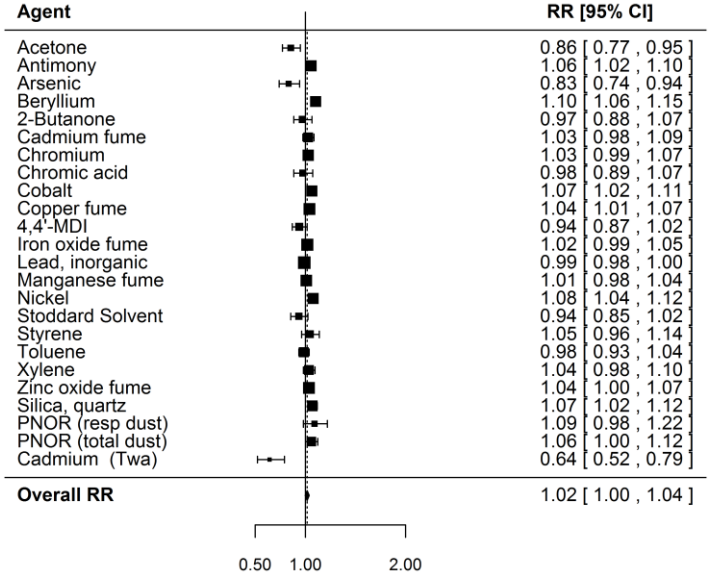
PEL citations in inspection=low (ref=none)



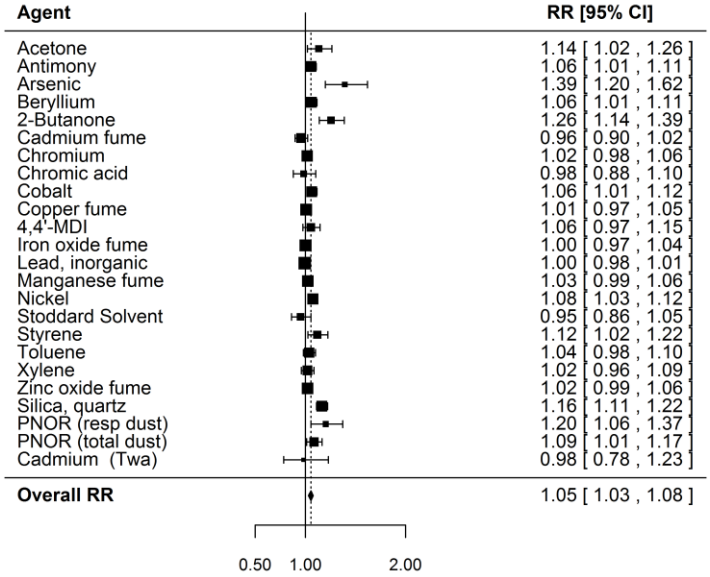
PEL citations in inspection=high (ref=none)



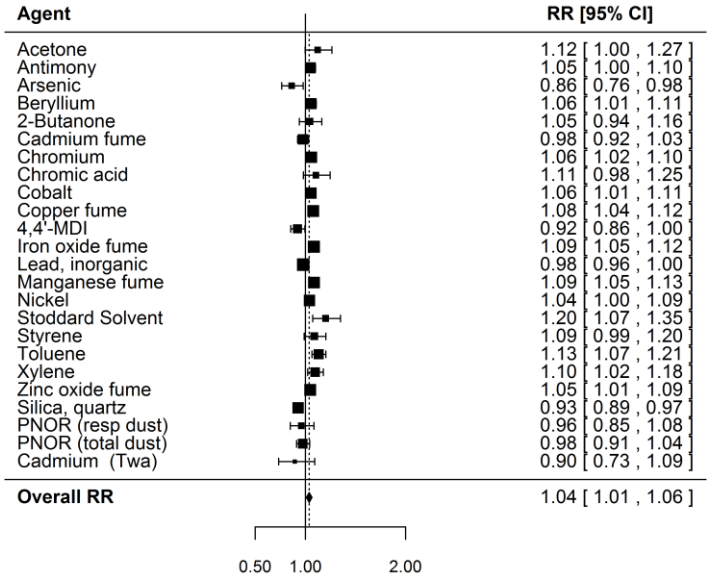
RespProt & HazComm citations in inspection=low (ref=none)



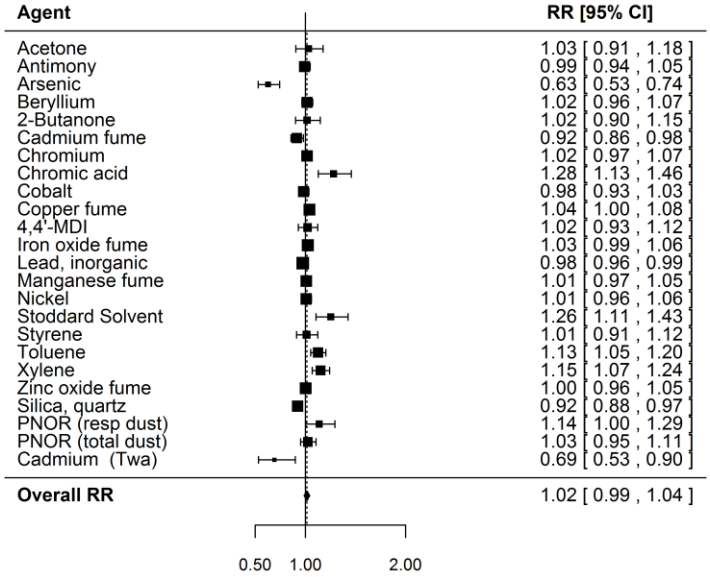
RespProt & HazComm citations in inspection=high (ref=none)



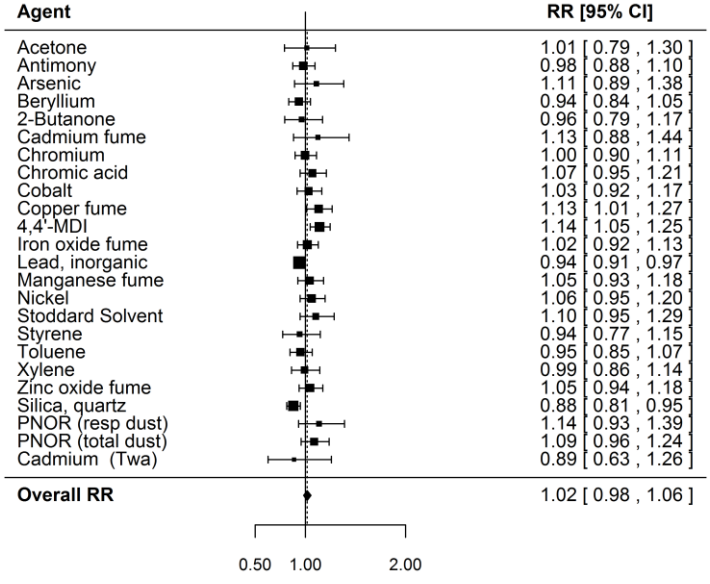
Other citations in inspection=low (ref=none)



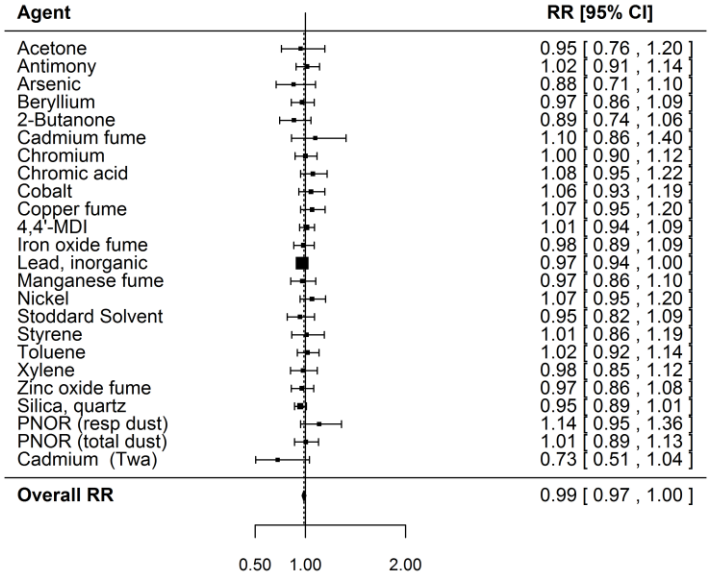
Other citations in inspection=high (ref=none)



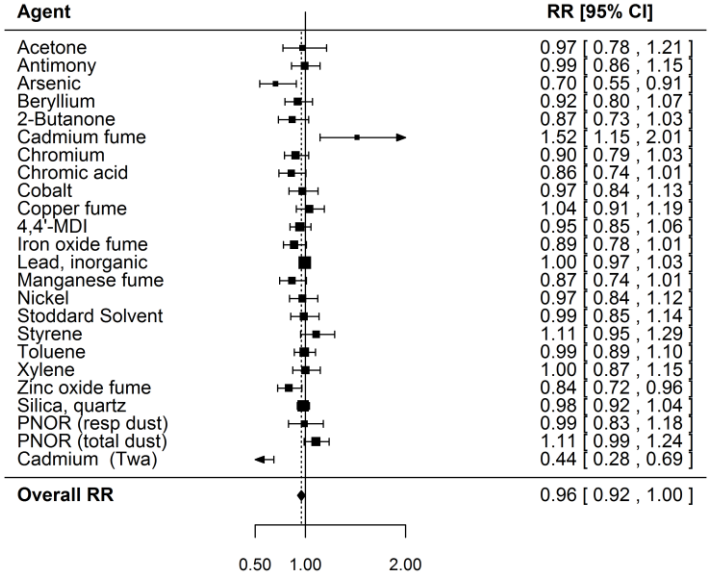
Other det_level in inspection=low (ref=none)



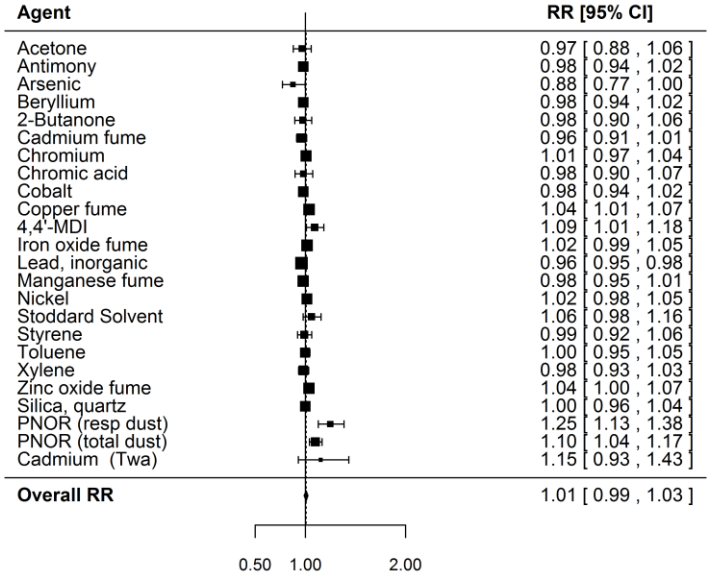
Other det_level in inspection=med (ref=none)



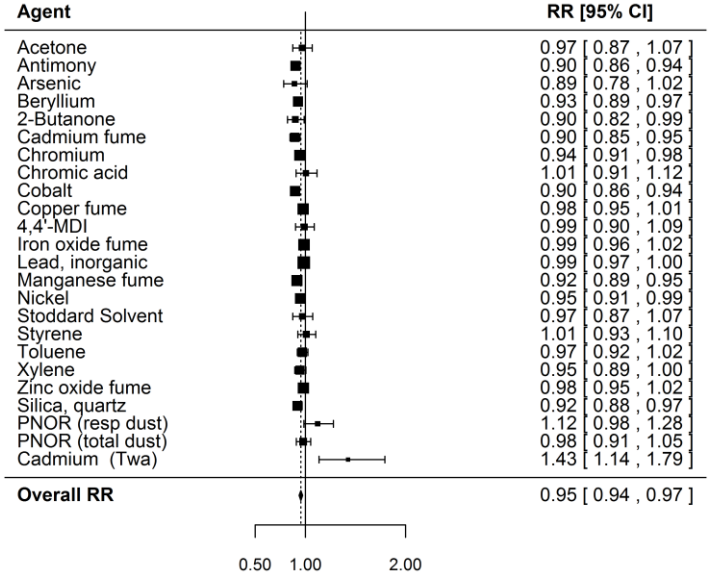
Other det_level in inspection=high (ref=none)



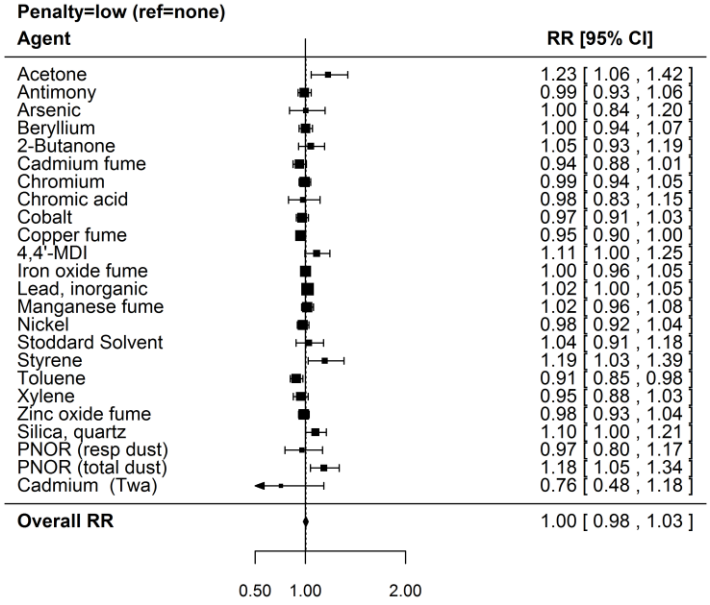
Establishment size=medium (ref=small)



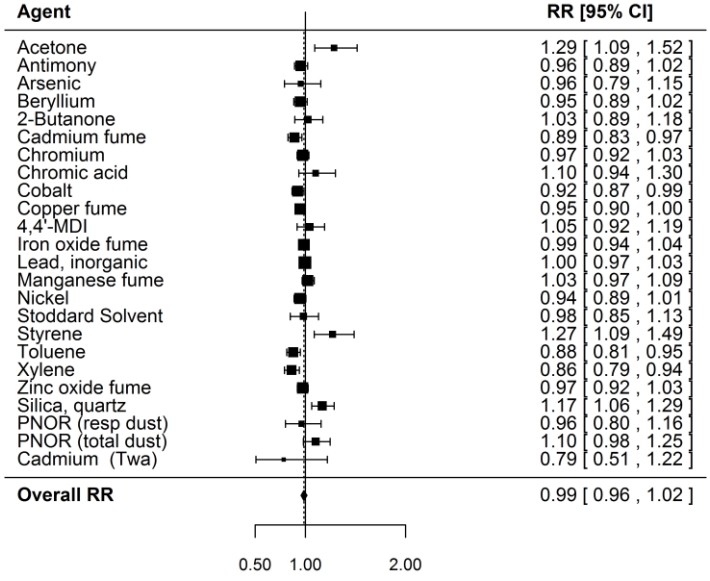
Establishment size=large (ref=small)



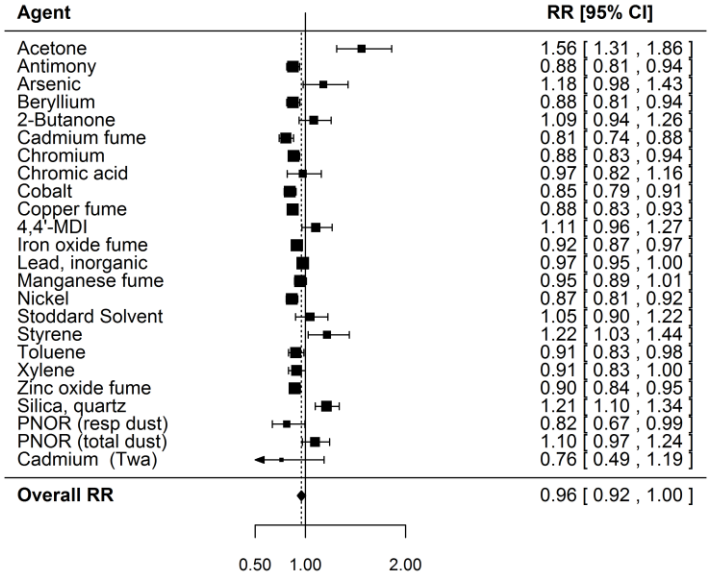
Penalty=low (ref=none)



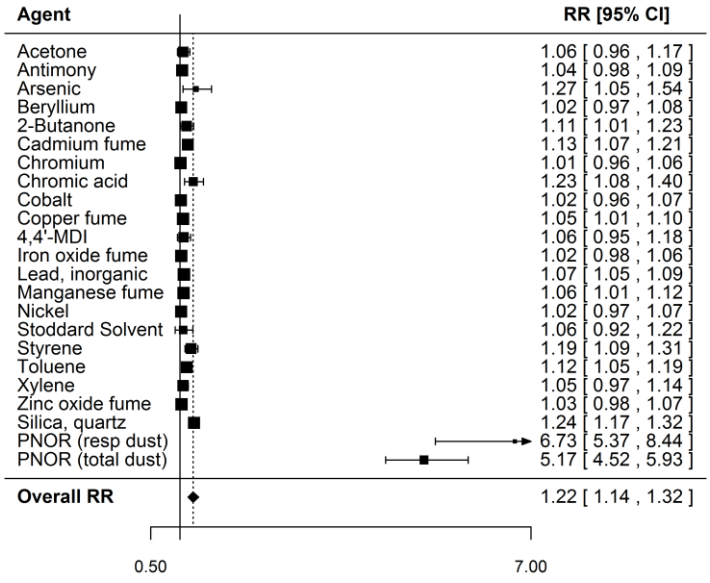
Penalty=medium (ref=none)



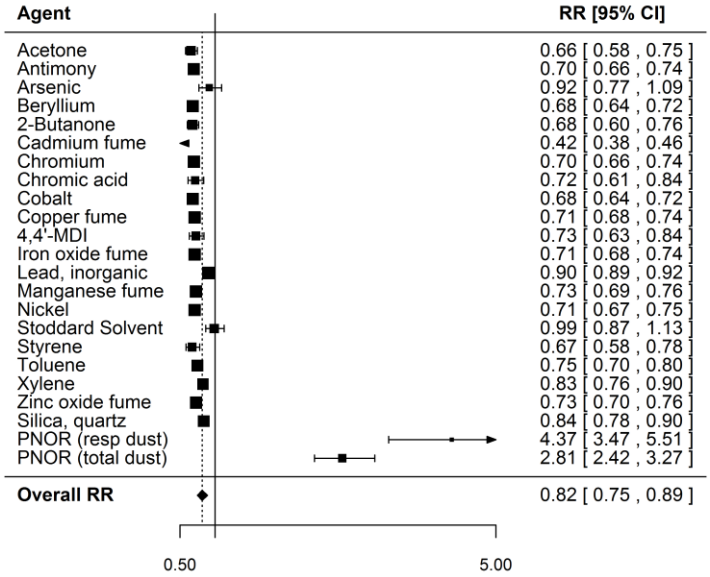
Penalty=high (ref=none)



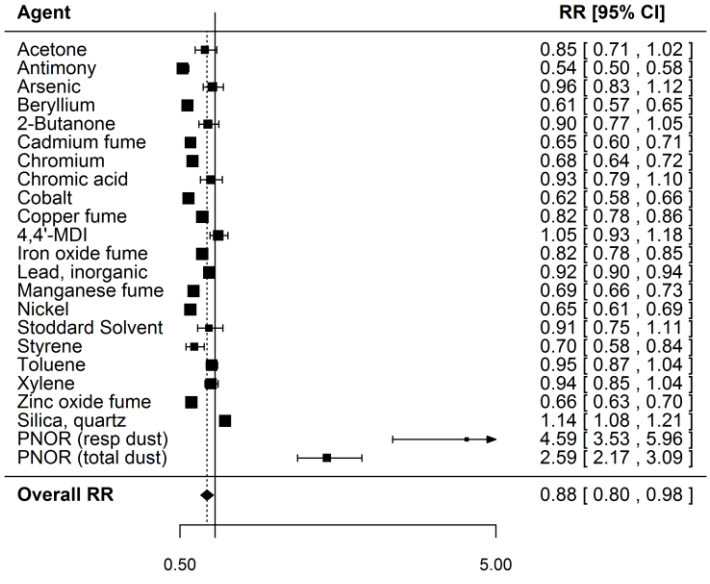
OSHA region=01_boston (ref=mean of OSHA regions)



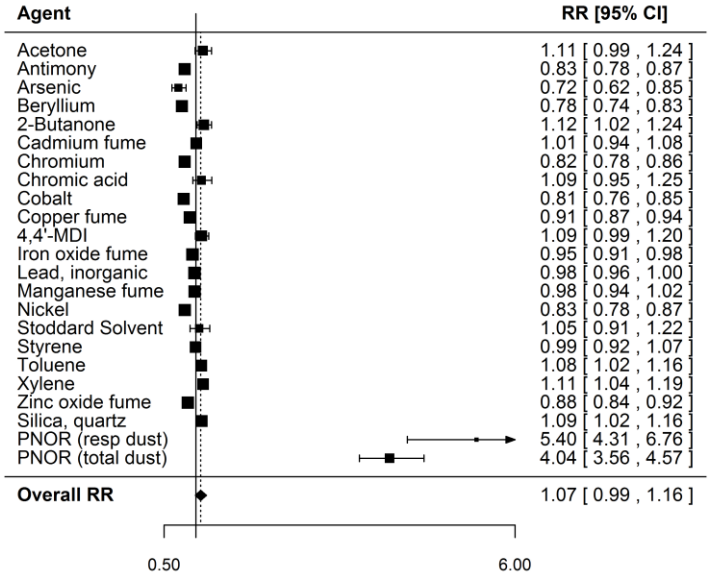
OSHA region=02_new_york (ref=mean of OSHA regions)



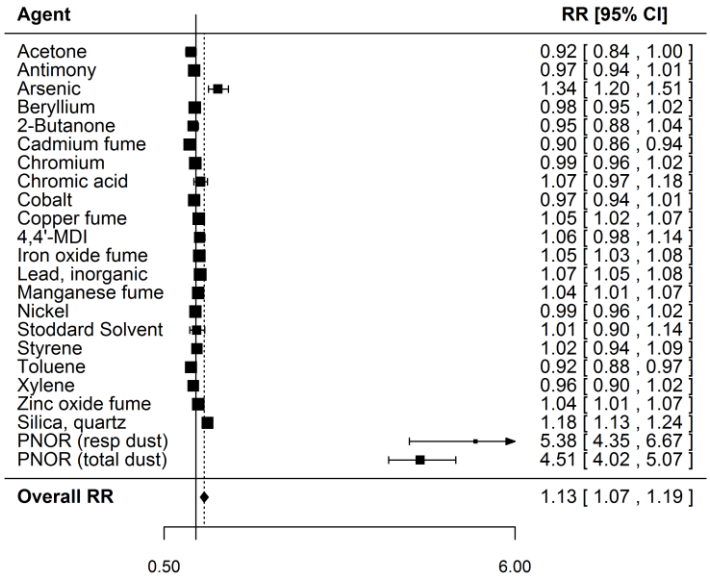
OSHA region=03_philadelphia (ref=mean of OSHA regions)

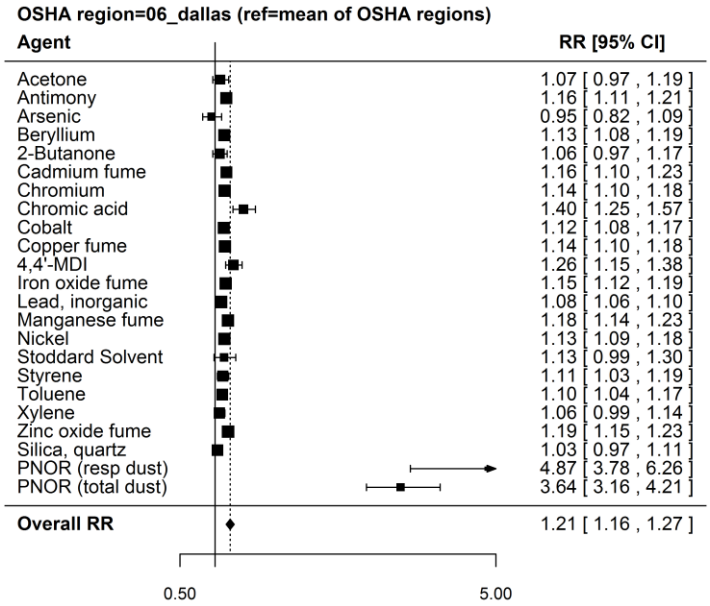


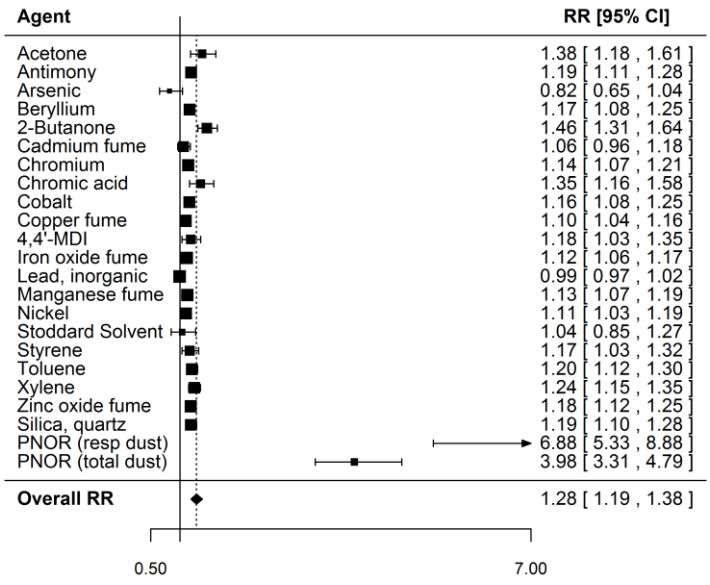
OSHA region=04_atlanta (ref=mean of OSHA regions)



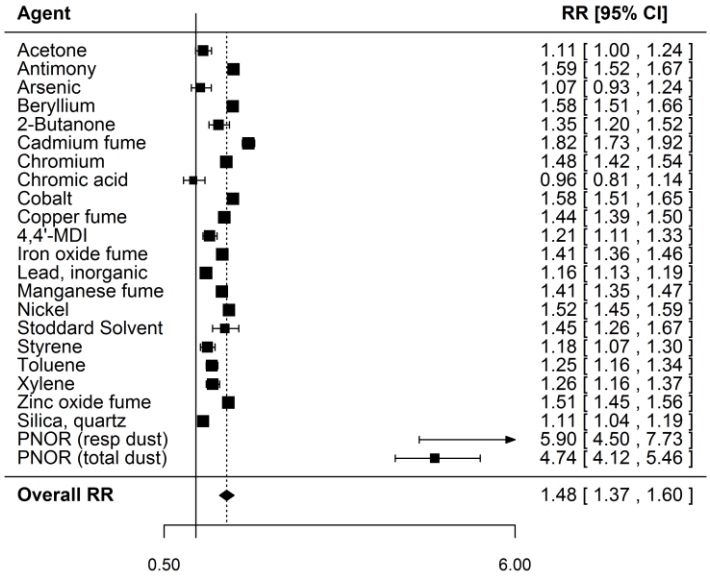
OSHA region=05_chicago (ref=mean of OSHA regions)



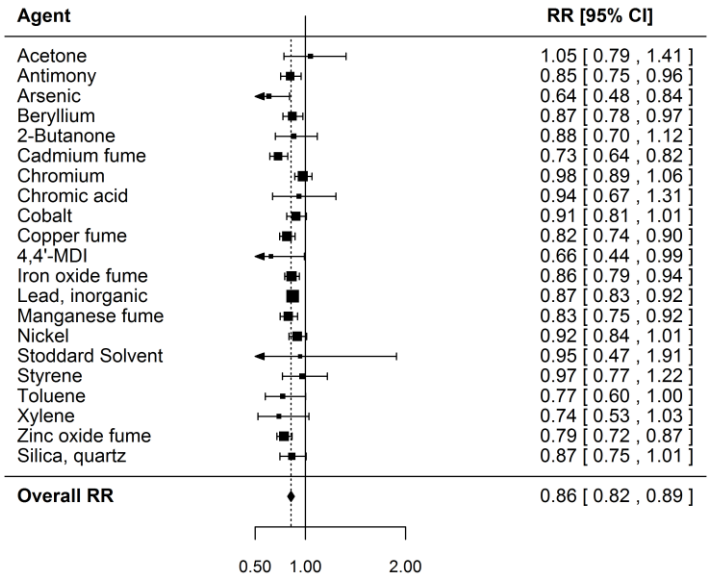

OSHA region=06_dallas (ref=mean of OSHA regions)
Agent
RR [95% CI]
Acetone 1.07 [0.97 , 1.19]
Antimony 1.16 [1.11 , 1.21]
Arsenic 0.95 [0.82 , 1.09]
Beryllium 1.13 [1.08 , 1.19]
2-Butanone 1.06 [0.97 , 1.17]
Cadmium fume 1.16 [1.10 , 1.23]
Chromium 1.14 [1.10 , 1.18]
Chromic acid 1.40 [1.25 , 1.57]
Cobalt 1.12 [1.08 , 1.17]
Copper fume 1.14 [1.10 , 1.18]
4,4'-MDI 1.26 [1.15 , 1.38]
Iron oxide fume 1.15 [1.12 , 1.19]
Lead, inorganic 1.08 [1.06 , 1.10]
Manganese fume 1.18 [1.14 , 1.23]
Nickel 1.13 [1.09 , 1.18]
Stoddard Solvent 1.13 [0.99 , 1.30]
Styrene 1.11 [1.03 , 1.19]
Toluene 1.10 [1.04 , 1.17]
Xylene 1.06 [0.99 , 1.14]
Zinc oxide fume 1.19 [1.15 , 1.23]
Silica, quartz 1.03 [0.97 , 1.11]
PNOR (resp dust) 4.87 [3.78 , 6.26]
PNOR (total dust) 3.64 [3.16 , 4.21]
Overall RR 1.21 [1.16 , 1.27]
0.50
5.00

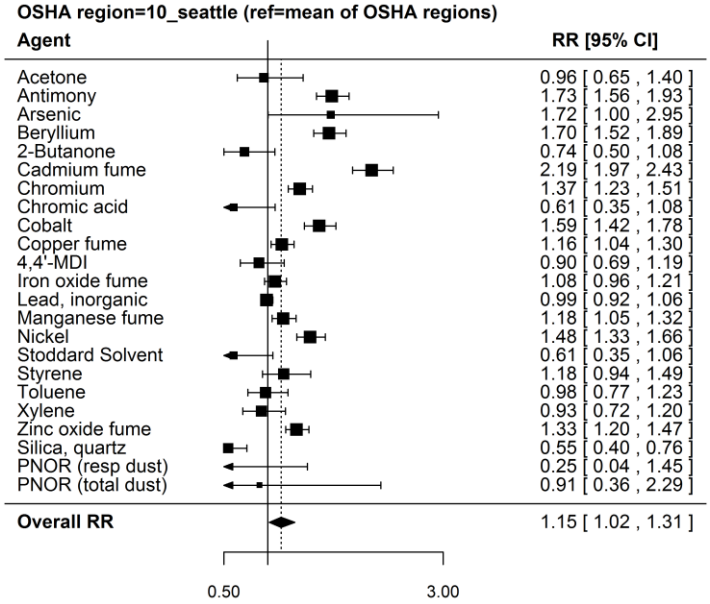

OSHA region=07_kansas_city (ref=mean of OSHA regions)
Agent
RR [95% CI]
Acetone 1.38 [1.18 , 1.61]
Antimony 1.19 [1.11 , 1.28]
Arsenic 0.82 [0.65 , 1.04]
Beryllium 1.17 [1.08 , 1.25]
2-Butanone 1.46 [1.31 , 1.64]
Cadmium fume 1.06 [0.96 , 1.18]
Chromium 1.14 [1.07 , 1.21]
Chromic acid 1.35 [1.16 , 1.58]
Cobalt 1.16 [1.08 , 1.25]
Copper fume 1.10 [1.04 , 1.16]
4,4'-MDI 1.18 [1.03 , 1.35]
Iron oxide fume 1.12 [1.06 , 1.17]
Lead, inorganic 0.99 [0.97 , 1.02]
Manganese fume 1.13 [1.07 , 1.19]
Nickel 1.11 [1.03 , 1.19]
Stoddard Solvent 1.04 [0.85 , 1.27]
Styrene 1.17 [1.03 , 1.32]
Toluene 1.20 [1.12 , 1.30]
Xylene 1.24 [1.15 , 1.35]
Zinc oxide fume 1.18 [1.12 , 1.25]
Silica, quartz 1.19 [1.10 , 1.28]
PNOR (resp dust) 6.88 [5.33 , 8.88]
PNOR (total dust) 3.98 [3.31 , 4.79]
Overall RR 1.28 [1.19 , 1.38]
0.50
7.00

OSHA region=08_denver (ref=mean of OSHA regions)



OSHA region=09_san_francisco (ref=mean of OSHA regions)




OSHA region=10_seattle (ref=mean of OSHA regions)
Agent
RR [95% CI]
Acetone 0.96 [0.65 , 1.40]
Antimony 1.73 [1.56 , 1.93]
Arsenic 1.72 [1.00 , 2.95]
Beryllium 1.70 [1.52 , 1.89]
2-Butanone 0.74 [0.50 , 1.08]
Cadmium fume 2.19 [1.97 , 2.43]
Chromium 1.37 [1.23 , 1.51]
Chromic acid 0.61 [0.35 , 1.08]
Cobalt 1.59 [1.42 , 1.78]
Copper fume 1.16 [1.04 , 1.30]
4,4'-MDI 0.90 [0.69 , 1.19]
Iron oxide fume 1.08 [0.96 , 1.21]
Lead, inorganic 0.99 [0.92 , 1.06]
Manganese fume 1.18 [1.05 , 1.32]
Nickel 1.48 [1.33 , 1.66]
Stoddard Solvent 0.61 [0.35 , 1.06]
Styrene 1.18 [0.94 , 1.49]
Toluene 0.98 [0.77 , 1.23]
Xylene 0.93 [0.72 , 1.20]
Zinc oxide fume 1.33 [1.20 , 1.47]
Silica, quartz 0.55 [0.40 , 0.76]
PNOR (resp dust) 0.25 [0.04 , 1.45]
PNOR (total dust) 0.91 [0.36 , 2.29]
Overall RR 1.15 [1.02 , 1.31]
0.50
3.00

CHAPITRE 5- Non-detects in OSHA's IMIS databank, are they short term or 8-hour shift-long samples? Prediction for 54 chemicals using recursive partitioning statistical methods

Non-detects in OSHA's IMIS databank, are they short term or 8-hour shift-long samples? Prediction for 54 chemicals using recursive partitioning statistical methods

Philippe Sarazin (1,6), Dan Vatnik (2), George Luta (3), Igor Burstyn (4), Laurel Kincl (5), Jérôme Lavoué (2,6)

(1) Institut de recherche Robert-Sauvé en santé et en sécurité du travail, Chemical and Biological Hazards Prevention, Montréal, Québec, Canada

(2) University of Montreal Hospital Research Centre, Montréal, Québec, Canada

(3) Georgetown University, Biostatistics, Bioinformatics, and Biomathematics, Washington, District of Columbia, United States

(4) Drexel University, Environmental and Occupational Health, Philadelphia, Pennsylvania, United States

(5) Oregon State University, College of Public Health and Human Sciences, Corvallis, Oregon, United States

(6) Université de Montréal, Department of Occupational and Environmental Health, Montréal, Québec, Canada

Corresponding Author:

Philippe Sarazin
Institut de recherche Robert-Sauvé en santé et en sécurité du travail
505, Boul. de Maisonneuve Ouest
Montréal, QC, Canada
H3A 3C2

Email: philippe.sarazin@irsst.qc.ca
Phone: (514) 288-1551 (Ext. 402)

Philippe Sarazin a contribué de façon majeure à la conception et au design de l'étude, à la mise en place du plan d'analyse des données, a été responsable de l'interprétation et de la présentation des résultats, de la rédaction du manuscrit et de l'édition de l'article.

5.1 Abstract

Objectives: The Integrated Management Information System (IMIS) is the largest multi-industry source of exposure measurements available in North America. However, the lack of information on the censoring value and sampling duration of non-detected (ND) measurements in IMIS prevents users from properly handling these results. The Occupational Safety and Health Administration (OSHA) released in 2010 a dataset of over 1 million sample results analysed at OSHA's central laboratory in Salt Lake City (Chemical Exposure Health Data (CEHD)). The CEHD data contains the analytical sample results and associated details of the measurement corresponding to records in IMIS. We undertook this study to gain insight into which ND results stored in IMIS are full-shift, and which are short-term samples, based on information available in CEHD.

Methods: The analyses were restricted to measurements from IMIS and CEHD taken between 1984 and 2009. Classification and regression trees (CART) models were applied to predict which ND results stored in IMIS are time-weighted average (TWA) or short-term samples, for 54 agents. Predictions were based on variables available in both IMIS and CEHD databanks (reason for inspection, scope of inspection, OSHA plan, OSHA region, union status, year of sampling, and industry).

Results: The median proportion of ND results in the datasets used to build CART models was 37% across agents (interquartile range (IQR) = (22%; 62%)). The industry was the most important variable in classifying ND results in TWA or short-term categories for 49 out of the 54 CART models, with a median relative importance of 52% across agents

(IQR = [44%; 59%]). CART models predicted more accurately which IMIS ND results were TWA or short-term samples compared to simple methods of assignment (e.g. assignment of the most frequent category from detected values) for the most relevant agents (i.e. with high proportions of ND, short-term, and TWA results). Predicted proportion of TWA for ND results in IMIS was at least 10% lower compared to detected results for all solvents, gases, and isocyanates, whereas it was similar for metals and dusts.

Conclusions: We developed a transparent and reproducible procedure that provides a systematic way to predict which ND results stored in IMIS are full-shift TWA or short-term samples, for 54 chemical agents. The model prediction results can be used to facilitate the assignment of a censoring value to ND results, thus helping to draw more accurate portraits of exposure levels in IMIS.

5.2 Introduction

Multi-industry occupational exposure databanks (OEDBs) have been used as sources of data for epidemiology or surveillance of the workplace (Stewart and Rice, 1990; Gomez, 1997; Rajan *et al.*, 1997; LaMontagne *et al.*, 2002; Lavoue *et al.*, 2008). Set up in several countries in the early 1980s, these databanks contain measurements made by governmental agencies for regulatory and prevention activities. Examples of such databanks include the U.S Integrated Management Information System (IMIS) (Stewart and Rice, 1990), the U.K. National Exposure Database (NEDB) (Burns and Beaumont, 1989), and COLCHIC in France (Vincent and Jeandel, 2001).

The IMIS databank, recently replaced by the OSHA Information System (OIS) (US Department of Labor, 2014), is the largest multi-industry source of exposure measurements available in North America. IMIS contains air sampling measurement results from surveys performed by OSHA inspectors to verify compliance to Permissible Exposure Limits (PELs). Along with the measured values, it contains variables potentially useful for occupational hazard surveillance such as the Standard Industrial Classification (SIC) codes, location and identity of the company inspected, and reason for inspection (Froines, 1989; Lavoue *et al.*, 2008; Sarazin *et al.*, 2016b). IMIS measurement data have been used in the past to evaluate historical occupational exposures to various physical and chemical agents (Coble *et al.*, 2001; Middendorf, 2004; Yassin *et al.*, 2005), identify factors associated with exposure levels (Gomez, 1997; Henn *et al.*, 2011; Sarazin *et al.*, 2016b), and in support of exposure assessment in epidemiology (Hamm and Burstyn, 2011; Lee *et al.*, 2015).

Even if IMIS represents a great potential source of information, results stored within this databank cannot be regarded, by default, as representative of the exposures experienced by the general U.S. working population. The targeting of workplaces for enforcement visits and the selection of workers for exposure monitoring are non-random and may overrepresent situations with higher- or lower-than-average exposures. Recent reports (Henn *et al.*, 2011; Lavoue *et al.*, 2013; Sarazin *et al.*, 2016b) investigating associations between exposure levels and ancillary variables in IMIS concluded that inspection selection processes and targeting practices influence to some extent the exposure levels recorded by OSHA. Handling of non-detected (ND) measurements represents an additional challenge in using the IMIS databank. The lack of information on the censoring value and sampling duration of ND results in IMIS complicates their interpretation.

As discussed in previous reports (Lavoue *et al.*, 2013; Sarazin *et al.*, 2016b), the status of a measurement coded as a ND is provided in the same variable that identifies a sample as a shift-long or short-term measurement. This implies that it is not possible to know whether a ND was measured as a shift-long time-weighted average (TWA) or a short-term value. This precludes users from properly handling the ND results because one does not know whether a non-detect was a full-shift TWA with lower limit of detection (LOD) or a short-term sample with higher LOD. This therefore prevents users from properly assigning a censoring level to ND records in IMIS.

Because of this limitation, different approaches have been adopted by investigators to handle non-detects in IMIS. ND results have usually been excluded from analysis in early studies based on IMIS data (Froines *et al.*, 1986; Froines *et al.*, 1990; Melville and Lippmann, 2001; Lurie and Wolfe, 2002). More recently, NDs have been included in analyses by expressing exposure levels as a simple dichotomy below/above the exposure limit instead of a continuous variable (Hamm and Burstyn, 2011; Henn *et al.*, 2011; Sarazin *et al.*, 2016b). Some attempts have also been made to evaluate the impact of excluding NDs from analysis (Lavoue *et al.*, 2008; Lavoue *et al.*, 2011) and found it was non-negligible.

OSHA released in 2010 a dataset of over 1 million sample results analysed at OSHA's central laboratory in Salt Lake City (Chemical Exposure Health Data (CEHD)) (Lavoue *et al.*, 2013). The CEHD data contains the analytical sample results and associated details corresponding to records in IMIS. This provides the opportunity to use the sampling time available for ND results in CEHD to gain insight into which ND results stored in IMIS are full-shift TWA, and which are short-term samples.

Our objective was to explore the possible prediction of type of exposure (i.e. TWA, short-term) of ND results in IMIS from ancillary information reflecting characteristics of the measurements, using recursive partitioning statistical methods.

5.3 Methods

The OSHA databanks: Integrated Management Information System (IMIS) and Chemical Exposure Health Data (CEHD)

OSHA maintains two separate databanks that include measurement results collected during compliance inspections conducted across the US between 1979 and 2010. The IMIS exposure databank was accessed through a Freedom of Information Act request. Each record included information about the company inspected, including the name and address of employer, and the total number of workers at the worksite. Date, sample number, sampled exposure level, type of sample (i.e. area, bulk, personal, screening), type of inspection (i.e. complaint, referral, follow-up, planned), and whether the inspection is conducted by federal OSHA or under an OSHA state plan are also recorded. The CEHD databank was accessed online (OSHA, 2015). It was made available in 2010, and contains the analytical results and associated details of the measurement collected by OSHA inspectors. The CEHD data supplements the IMIS data with the sampling duration, analytical method, and presence of other substances on the same sampling media. OSHA officers perform calculations using the sample results (e.g. a time-weighted average concentration, TWA, calculated from several sequential samples on a worker), and record the result of their calculation in IMIS. The CEHD databank is only partially overlapping IMIS since records in CEHD are mainly samples collected by federal OSHA and analysed at OSHA's central laboratory in Salt Lake City.

Data preparation

The IMIS and CEHD extracts available for this analysis contained data from the time periods 1979–2011 and 1984–2009, respectively. The analyses were therefore restricted to measurements from both databanks taken between 1984 and 2009. The IMIS extract contained 851,987 records corresponding to 107,647 inspections, covering 1,054 agents. Cleaning of IMIS data was described in Sarazin *et al.* (2016b). Briefly, records corresponding to area, bulk, blood, urine, wipe, screening samples, noise, and exact duplicate samples were excluded. The CEHD online dataset contained 1,908,373 records corresponding to 1,082 agents. Cleaning of CEHD data was described in detail in appendix 1 in Lavoue *et al.* (2013). Briefly, records were removed if they were not personal measurements, irrelevant for exposure assessment (e.g. blank samples), had uninterpretable misspellings, missing information, or judged erroneous. The analysis was restricted to all chemical agents analysed in Sarazin *et al.* (2016b) since the current study originally aimed at improving treatment of ND results in the published analysis.

Linkage between IMIS and CEHD

Both datasets were combined using the 'sampling number' variable which was considered as identifying a unique 'evaluation' made by an inspector (e.g. a worker's full shift). Hence several samples in CEHD for one sampling number would correspond to partial shift measurements aggregated by the inspector prior to recording in IMIS through the calculation of a time weighted average. 'Sampling number' is not a perfect variable for linking CEHD to IMIS, since in some cases aggregated sampling time seems

unrealistically high (i.e. >600 minutes), and some records in IMIS have the same sampling number. However, preliminary analyses on three chemical agents (lead, formaldehyde, and toluene) showed that these issues were relatively infrequent (i.e. <5%). Multiple records tied to a single 'sampling number' in CEHD were therefore treated as sequential partial-shift measurements and aggregated to calculate total sampling time for the evaluation. The aggregated value was flagged as a non-detect when all partial-shift measurements were reported as non-detects. This approach to flag non-detects yielded a discordance of 2.7% between the status of measurements for the records that were overlapping between the IMIS and CEHD datasets (i.e. were ND in IMIS but flagged as detected in CEHD). In cases where measurements were discordant we defaulted to the IMIS status recorded by the inspector since there was no reason to question his or her choice.

Development of prediction models for type of exposure (TWA, short-term)

The focus of the current study was to use information in IMIS and CEHD in the overlapping dataset to predict which ND results in the non-overlapping IMIS dataset are TWA or short-term samples.

CART models

Classification and regression trees (CART) models (Berk, 2008; Strobl *et al.*, 2009), recently applied in occupational health and epidemiology studies (Friesen *et al.*, 2013;

Wheeler *et al.*, 2013; Van Hulst *et al.*, 2015), were used since they allow modeling nonlinear associations between predictor variables and are easily interpretable. Tree-based approaches predict decisions based on a sequential splitting pattern that resembles an upside-down tree, with the 'root' at the top below which are nodes that divide observations into branches (Berk, 2008). In our case, the method produces a classification tree by dividing IMIS ND results into TWA and short-term subgroups based on a number of predictor variables. The tree-building process starts by considering the set of predictor variables and selects the variable that produces two subsets of ND results with the greatest purity (i.e. where ND results within each subset are most alike in terms of the outcome – TWA or short-term). Two factors are considered when splitting data into its daughter nodes: the goodness-of-split (Leblanc and Crowley, 1993) and the amount of impurity in the daughter nodes (Berk, 2008). The splitting process is repeated until the model meets specified stopping criteria (e.g. complexity parameter set to control the growth of the tree, minimum number of observations) and terminal nodes have been reached. At each terminal node is the outcome prediction for the specific subset of the data. For example, the predicted outcome in the terminal node will be 'TWA' if the number of ND results belonging to the 'TWA' class in the node exceeds the pre-specified threshold (e.g. 50%). CART models do not make any distributional assumption and treat the data generation process as unknown. They also do not assume additivity of the predictors which allows them to identify complex interactions.

Ancillary information examined

Ancillary information common to both databanks and judged potentially useful for predicting type of exposure of IMIS ND results were included in analysis. Type of inspection is a record of the reason for conducting each inspection. Only five of the twelve possible categories – planned, complaint, referral by a safety compliance officer or other source, follow-up, and monitoring – were retained for analysis since they represented more than 95% of the data. The 'follow-up' and 'monitoring' categories were combined since they both correspond to inspections where the compliance officer returned in a particular facility. For the variable federal/state OSHA plan, CEHD sampling results were categorized as Federal or State according to the states' status at the time of sampling. For the SIC variable, a partial aggregation of data across the 4-digit SIC categories was performed so that ~30 industry groups were obtained for each chemical agent in the analysis. This yielded a manageable number of categories and still allowed variation across a significant number of groups. For each agent, when fewer measurement results than a predetermined cut point was available for a 4-digit SIC category, the more specific digit was dropped to create a broader category. Cut points were determined by dividing the total number of measurements for the specific agent by 60. The cut points varied from 15 for the least frequent agent to 949 for the most frequent agent. The process was repeated until the number of measurements in the category was greater or equal to the cut point or the code was reduced to a SIC major division (1-digit). Finally, if a 1-digit code was not associated with more measurements than the cut point, it was put into an 'other' category. This system of classifying SIC codes is typical of those employed when working with other similar analyses (Lavoue *et al.*, 2008; Lee *et al.*,

2015; Sarazin *et al.*, 2016b). Scope of inspection, union status, year of inspection, and OSHA region were also included in analysis (Table I).

Modeling datasets

For each chemical agent, the merged IMIS and CEHD datasets were restricted to ND results and split into three datasets: records that were in common to both IMIS and CEHD datasets ('overlapping' dataset), records that were only in CEHD ('CEHD-only') and records that were only in IMIS ('IMIS-only') (Figure 1). Based on the sampling time information in CEHD, we identified the exposure type of ND results as either TWA or short-term in the 'overlapping' and 'CEHD-only' datasets (records with sampling time >60 minutes were identified as TWA). To get insight about the adequacy of this criterion, we used the 'overlapping' detected dataset where TWA and short-term status can be compared between IMIS and CEHD. There was 6.1% discordance between the TWA status (i.e. were TWA in IMIS but identified as short-term in CEHD) and 23.2% discordance between the short-term status (i.e. were short-term in IMIS but identified as TWA in CEHD).

The 'overlapping' datasets were used to build the models and the 'CEHD-only' datasets were used to perform external validation of the models. The final models were then used to predict the exposure type of ND results in the whole IMIS dataset.

Metrics used to assess CART model performance: overall error from confusion table

A confusion table is a table that is used to describe the performance of a classification model (such as CART) on a set of test data for which the true values are known. It is a 2 X 2 table that contains information about the observed classes and the classes that the model assigns which allows investigating misclassification rates by group. The cross-tabulations of the predicted and observed type of exposure (i.e. TWA, short-term) of ND results allowed investigation of error rates from the CART models. A mock example of a confusion table for a given agent is presented in Table II. Columns of the table represent the type of exposure that the model predicted for the ND results, and rows represent the true type of exposure of the ND results. Thus, the diagonal of the table represents the numbers of ND results correctly classified by the model, while incorrectly classified results are represented by the numbers in other cells. In this example, there were 25 ND records whose type of exposure was short-term, but the model predicted that they would be in the TWA group. Similar interpretations can be made for the rest of the cells in the table. The last column in Table II specifies the error rate of the model broken down by outcome categories and the overall error rate of the model. The overall error rate is the metric that was used throughout the study since we had no basis to attribute a higher cost to incorrect classification of either the TWA or short-term categories.

Subset of agents to select CART parameters

To generate CART models, two parameters needed to be selected/optimised, including minimum number of records required in a node in order to be split (minimum split), and the splitting type (Gini or Information). Multiple combinations of parameter values were tested on a subset of agents (acetone, butoxyethanol, cadmium dust, formaldehyde, methyl chloroform, phenol, and toluene). These agents were selected in order to cover a wide range of proportions of TWA among ND results (from 20% to 76%). We tested three minima for the number of observations per node in order for a split to be attempted (2%, 4% and 8% of the agent's total observations), and two indicators of node purity (Gini index and Information Gain criteria (Raileanu and Stoffel, 2004)), which reaches its minimum for perfectly pure nodes and its maximum when observations are distributed evenly between classes at a given node.

Final CART models for all agents

Based on the tested combinations of CART model parameters for the 7 agents, the minimum number of records to allow a split within a node was set to 2% of the sample size, and the indicator of node purity was set to Information Gain criteria.

In the final analysis across all agents, with the previous parameters fixed, the full range of the complexity parameters (CP), which controls how large a tree can grow (from 0.1-simplest model to 0.0001-most complex model), were tested for each chemical agent. The ten-fold validation technique (Strobl *et al.*, 2009; Wheeler *et al.*, 2013; Ripley, 2015) was used in order to choose the optimal CP value. This technique is used to prevent the tree

from overfitting the data, where the best tree is based on a rule in which the overall error rate is no more than 1 standard error (SE) larger than the best tree.

Evaluation of CART model performance

Internal performance

For each chemical agent, bagging, which stands for 'Bootstrap Aggregation' (Berk, 2008; Strobl *et al.*, 2009), was used to assess the internal performance of the CART models built from the 'overlapping' datasets. Bagging is a simulation technique where 95% of the data is randomly selected with replacement to build a regression tree, while the remaining 5% of data are used to validate the tree by comparing the predicted to the real results (1000 iterations were performed). We computed the median overall error of the 1000 iterations.

Comparison to two simpler methods of assignment

For each chemical agent, the performance of CART models was compared to two simpler methods of assignment of TWA and short-term which do not involve modeling: random assignment and assignment of most frequent category. For 'random assignment', ND records were assigned category TWA or short-term based on their respective proportions in the detected dataset. For 'assignment of most frequent category', ND records were set

all as TWA or all as short-term based on the most frequent category in the detected dataset.

External performance ('CEHD-only')

For each chemical agent, measures of the predictive performance of the CART model were evaluated on the validation dataset 'CEHD-only'. We evaluated the agreement of the model predictions with the actual type of exposure of ND results (identified as TWA or short-term) in the validation set using the overall error from the confusion table.

Application of predictions in a previously published analysis

The predictions of our CART models were applied to the study of Sarazin *et al.* (2016b), which evaluated the relationship between exposure levels in IMIS and ancillary variables reflecting the characteristics of establishments selected for inspection and OSHA sampling practices. The same modeling approach as in Sarazin *et al.* (2016b) was used with the same set of predictors, but this time incorporating information on which ND results are TWA or short-term samples, therefore allowing to include the exposure type variable in the logistic models. Models were applied to three frequently measured agents (2-butanone, formaldehyde, and xylene) with >15% of NDs and >10% of both TWA and short-term among their detected results. Effects were estimated for each predictor and compared to previously published results.

Software

All analyses were performed using the R 3.1.3 statistical software (R Development Core Team, Vienna, Austria), with the package rpart (Ripley, 2015) for recursive partitioning, and mgcv (Wood, 2014) for binomial GAM modeling.

5.4 Results

Descriptive analysis

The initial CEHD dataset contained 1,034,000 analytical measurements, which were reduced to 588,818 average concentration results corresponding to 36,442 inspection visits for the period 1984 to 2009. The IMIS dataset contained 511,047 measurements from 66,827 inspection visits for the period 1979 to 2011.

Sixty-three of the 77 chemical agents analysed in the IMIS study of Sarazin *et al.* (2016b) were selected for analysis (21 organic solvents, 21 metals and their compounds, 6 gases, 6 dusts/fibers, 4 isocyanates, and 5 other agents) (Table III), as the 14 others were not part of the original CEHD cleaning procedure (Lavoue *et al.*, 2013). The median proportion of ND results in the 'overlapping' datasets was 37% across agents [interquartile range (IQR) = (22%; 62%)], with a maximum of 93% for antimony and silica cristobalite. The proportion of TWA results in the detected 'overlapping' datasets ranged from 4% (4,4'-MDI) to 100% (silica cristobalite, fluoride, sulfuric acid, ethyl acetate, respirable dust), with a median of 80%.

In order to assess the performance of CART models, we stratified our analysis into 3 groups: agents with a high proportion of ND results in the 'overlapping' dataset and a high proportion of both TWA and short-term in the detected 'overlapping' dataset (group 1: >15% NDs and >10% of both TWA and short-term), agents with a high proportion of ND results but with a majority of either TWA or short-term (group 2: >15% NDs and

>90% of either TWA or short-term), and agents with a low proportion of ND results (group 3: <15% NDs). We consider that predicting accurately which ND results are TWA or short-term is more important for group 1 agents since their proportion of NDs is high and there is no indication that a specific type of exposure is most frequent for these agents.

When comparing the 'overlapping' ND (used to build CART models) and detected datasets (Table III), the proportion of TWA was lower for ND results for 24 of the 29 group 1 agents (median difference of 23%). For group 2 agents (N=27), the proportion of TWA results was also lower for ND compared to detected for all solvents (7/7; median difference of 69%), while it was similar for all metals and isocyanates (13/13; median difference of 8%). When comparing the 'overlapping' ND dataset to the 'CEHD-only' ND dataset (used for external validation), the proportion of TWA results was lower in 'overlapping' for 62 of the 63 agents (median difference of 18%).

Nine out of the 63 chemical agents selected had to be removed during the analyses since their number of 'overlapping' ND results used for CART model building was too low (n<110) and prevented success of the model building process (copper dusts, ethyl acetate, mercury, PNOR-respirable dusts, PNOR-total dusts, silver, sulfur dioxide, tin, and trichloroethylene).

Variable importance in final CART models

Table IV summarizes the importance of predictor variables in classifying which ND results are TWA or short-term in final CART models across the 54 chemical agents. The variable importance plot for manganese fume is illustrated in Figure 2, showing that SIC code was the most important variable for this agent. The SIC code was the most important variable in classifying whcih ND results are TWA or short-term in 49 out of the 54 CART models, with a median relative importance of 52% across agents (IQR = [44%; 59%]). OSHA region and year of sampling followed with median relative importance of 16% and 14%, respectively.

CART models performance

The performance of CART models in classifying which ND results are TWA or short-term for the 54 agents was assessed using the overall error computed from confusion tables.

Internal performance: comparison with simpler methods of assignment of TWA and short-term

The internal performance was assessed on the 'overlapping' ND results used for CART model building using the bagging approach (Table V). The median CART model bagging overall error (i.e., proportion of ND results incorrectly classified by the model) was higher for group 1 agents (0.225, range=[0.101;0.325]) compared to group 2 agents (0.123, range=[0.064;0.339]). The CART model bagging overall error was at least 10%

lower than 'random assignment' error for 42 out of 54 agents. Eleven of the 12 agents that showed similar or better performance of the 'random assignment' method were in group 2 (>90% of either TWA or short-term in their detected dataset), while the other agent was in group 3 (<15% NDs). The bagging overall error was at least 10% lower than 'assignment of most frequent category' error for 34 out of 54 agents. Sixteen out of the 20 agents that showed similar or better performance of the 'assignment of most frequent category' method were in group 2 or 3.

External performance (CEHD-only)

The external performance was assessed by predicting which ND results in the 'CEHD-only' dataset are TWA or short-term (Table V). The overall prediction error of CART models was at least 10% higher than the internal overall error for 37 out of 54 agents. The median overall prediction error of CART models was higher for group 1 agents (0.430, range=[0.062;0.695]) compared to group 2 agents (0.195, range=[0.043;0.707]).

CART models predictions

For each agent, CART models were used to predict which ND results in IMIS are TWA or short-term. Number of ND results in IMIS, and predicted proportions of TWA are presented for each agent in Table VI. Predicted proportion of TWA for ND results in IMIS was at least 10% lower compared to detected results for 32 out of 54 agents (all 32

agents were solvents, gases, or isocyanates). Nineteen out of the 22 agents with similar proportions of TWA in IMIS ND and detected results were metals or dusts.

Application of predictions in the previously published analysis

Table VII shows the influence of predictor variables on the odds of exposure exceeding TLV in IMIS for 2-butanone, formaldehyde, and xylene, incorporating information on which ND results are TWA or short-term samples. The six variables that showed an association with exposure levels for the 3 agents were the same as in Sarazin *et al.* (2016b) (inspection type, OSHA plan, OSHA region, amount of penalty, union status, and establishment size), and the amplitudes of the effects were similar. Exposure levels for shift-long TWA samples were close to two times lower than short-term samples for 2-butanone (OR = 0.51, 95% confidence interval (CI): 0.35–0.75) and formaldehyde (OR = 0.59, 95% CI: 0.53–0.67), and close to three times lower for xylene (OR = 0.33, 95% CI: 0.22–0.49).

5.5 Discussion

The interpretation of ND results in IMIS data is important given the high percentage of recorded non-detects (35%). The lack of knowledge on which ND results stored in IMIS are full-shift TWA or short-term samples precludes users from properly handling these measurements. We used the information on sampling time available in the CEHD databank to gain insight into which ND results stored in IMIS are TWA or short-term. Recursive partitioning methods were applied to predict the type of exposure (i.e. TWA, short-term) of non-detected sample results in IMIS. Predictions for 54 chemical agents for the period 1984-2009 were based on common variables available in IMIS and CEHD databanks. Using the CART model predictions, we investigated for three agents the associations between exposure levels and ancillary variables in IMIS including the exposure type variable, which allowed us to verify that excluding information on this variable did not confound our previous analyses (Sarazin *et al.*, 2016b).

CART models performance

Our CART models' predictive abilities indicate better performance for agents with a majority of either TWA or short-term measurements (i.e. group 2) compared to agents where measurement results are more evenly distributed between TWA and short-term categories (i.e. group 1). The error rates found in validation steps were around 10% for group 2 agents (i.e. high proportions of ND, and majority of either short-term or TWA), meaning that the CART model would incorrectly classify type of exposure around 10% of the time. The error rates for group 1 agents were around 20-25%. These higher error

rates for group 1 agents are disappointing since predicting accurately which NDs are TWA or short-term is more important for these agents. However, our results indicate that CART models performance for group 1 agents was still much better than the two other methods usually considered when handling ND results in IMIS, i.e. random assignment (average error rate = 53%) and assignment of most frequent category (average error rate = 56%). One could assume that an agent associated only with a TWA-PEL would have only TWA measurements, and therefore all ND results would be TWA for that agent. However, we noticed in our analysis that even if an agent was only associated with a TWA-PEL, the proportion of short term results for detected measurements could be substantial (e.g. >30%). This prevented usage of this simpler method of assignment of TWA and short-term which does not involve modeling. For group 2 agents, attributing the most frequent type of exposure (i.e. TWA, short-term) found in detected measurements to all ND results could be considered an approach just as valid as using our CART models. Finally, given the low proportion of ND results for group 3 agents, excluding ND results altogether would probably have little effect on analyses performed with these agents.

CART models predictions

The CART models' predictions for the 54 chemical agents (Table VI) in IMIS showed that proportion of TWA is generally lower for ND results compared to detected results for solvents and gases, which is expected since, with the same analytical limit of detection in mass, longer samples have a lower LOD in air concentration due to higher

sampling volume. This supports the hypothesis that we cannot use the proportions seen in the detected dataset to assign TWA or short-term to ND records since these data have different properties. It is possible that part of the difference seen between predicted proportions of TWA in ND and proportions of TWA in detected datasets may have been amplified by the tendency of CART models to overly predict TWA ND results as short-term. Indeed, for a majority of agents with higher error rates (i.e. solvents and gases), the rate of misclassification was much higher for TWA ND results (i.e. TWA wrongly classified as short-term) compared to short-term ND results. There are substantial differences between TWA and short-term error rates because the default settings in the rpart package (Speiser *et al.*, 2014; Ripley, 2015) seeks to minimize the overall error rate in internal validation steps, such as bagging. Therefore, the rate of misclassification will be lower for the most frequent category (i.e. short-term for solvents and gases), and higher for the least frequent category (i.e. TWA for solvents and gases). Weights could be applied to the model to decrease or increase the error rates of specific classes, if deemed more important that certain outcome groups would be accurately predicted. For the prediction of IMIS ND type of exposure, class weights were not applied to the model because nothing suggested that the cost of incorrect classification for each group is any different. Finally, this phenomenon probably only had a minor effect since the predicted difference between detected and ND results in IMIS data was also seen in the 'overlapping' dataset used for CART model building, where ND type of exposure is known.

Application of predictions in the previously published analysis

For the 3 tested chemical agents (2-butanone, formaldehyde, and xylene), controlling for type of exposure did not change significantly the associations between exposure levels and variables reflecting OSHA sampling practices. The same six variables had ORs away from the null for at least one category (inspection type, OSHA plan, OSHA region, amount of penalty, union status, and establishment size), which supports the conclusions presented in Sarazin *et al.* (2016b). Moreover, short-term measurements were approximately twice as likely to have a sample result exceed the TLV compared to TWA measurements, a result compatible with what has been obtained in analyses restricted to detected concentrations (Sarazin *et al.*, 2016b), where detected short-term concentrations were in average 3 times as high as TWA concentrations.

Limitations

The 'CEHD-only' dataset used for external validation of CART models' predictive abilities cannot be considered as a perfect validation dataset. Previous reports (Lavoue *et al.*, 2013; Sarazin *et al.*, 2016a) have indicated that there is a selection made by the OSHA officer of which sample results in CEHD are recorded in IMIS, and that the decision seems to depend on several factors, such as the level of exposure of the sample record. Therefore, the results that have not been recorded in IMIS (i.e. 'CEHD-only' dataset) might not be a valid gold standard for what has been recorded (i.e. 'overlapping' dataset).

Another limitation of our work is the lack of measurement level predictors since all the predictors considered are at the inspection level, i.e. predictors that take the same value for all the measurements within a specific visit. As a direct consequence all the ND observations from the same inspection will be classified as either TWA or short-term exposure measurements. It is unlikely that an inspector would only take measurements of one type, especially for agents having both TWA and short-term PELs. The creation of standardized occupation code from the crude job titles provided in IMIS (Burstyn *et al.*, 2014; Russ *et al.*, 2014) might allow finer predictions.

A third limitation of the present study relates to the equations used to regroup sequential CEHD samples to calculate the total sampling time for the evaluation (the 'sampling number' variable was used as a single identifier of an evaluation made by the inspector). Hence 23.2% of IMIS short-term samples in the 'overlapping' detected dataset were identified as TWA based on sample duration in CEHD. This suggests that CEHD samples linked to one sampling number could correspond to the monitoring of different situations. For example, it is possible that two 30 min. samples with the same sampling number are present in CEHD (therefore identified as TWA by our method), while the record in IMIS for the same sampling number is labelled short-term. Moreover, the discordance would have been even higher had we taken a threshold closer to the expected sampling time of a short-term sample to identify an aggregated CEHD evaluation as TWA. Indeed, we chose 60 min. as a threshold whereas short-term samples in IMIS, corresponding to peak, ceiling, or STEL limits, should theoretically be shorter than 30 min. These observations suggest that information available to link CEHD and IMIS is not perfect, which may have

partly confounded the attribution of TWA or short-term to ND results in the 'overlapping' (used for model building) and 'CEHD-only' datasets (used for external validation).

Finally, it is likely that knowing the particular circumstances under which ND samples are collected could allow one to make a judgement on whether a result is likely TWA or short-term (e.g. specific process associated with short tasks for an agent). Agent and a rather coarse industry code were the only information available in the CEHD databank, and both were indeed predictors of ND sampling duration. However, it would not have been possible to include finer information (e.g. studying each agent/industry combination to make a judgment about the most likely type of measurement, when possible) considering the wide range of agents analysed and that the whole industry spectrum was included.

Conclusions

Recursive partitioning is a valuable data exploration method in the study of chemical sampling results stored in exposure databanks. We developed a transparent and reproducible procedure that provides a systematic way to predict which IMIS ND results are TWA or short-term samples for 54 chemical agents based on common variables between IMIS and CEHD databanks. Findings showed better performance of our CART models compared to random assignment or assignment of most frequent category approaches for the most relevant group of agents (i.e. with high proportions of ND,

TWA, and short-term results). The model prediction results can be used to facilitate the assignment of a censoring value to ND results, thus helping to draw more accurate portraits of exposure levels in IMIS for a pollutant or an industry/occupation. Future methodological research is needed to better interpret ND results in OSHA's databanks.

5.6 Tables and figures

Table I : Variables used for CART model building

Variable	Description	Type	Number of samples (%)
Inspection type	Reason for conducting inspection	Nominal (4 categories)	
		(8) Planned	32,996 (28)
		(9) Complaint	62,320 (52)
		(10)Follow-up	6,309 (5)
		(11)Referral	17,222 (14)
Inspection scope	Comprehensive or partial survey of the establishment	Nominal (2 categories)	
		(5) Comprehensive	47,982 (40)
		(6) Partial	70,865 (60)
Union status	Union is present or not in the company monitored	Dichotomous	
		(3) Yes	41,233 (35)
		(4) No	77,614 (65)
OSHA plan	Inspection conducted by federal OSHA or under an OSHA state plan	Dichotomous	
		(3) Federal	117,393 (99)
		(4) State	1,454 (1)
OSHA region[a]	Identifies the OSHA region where the inspection took place	Nominal (10 categories)	
		(21)01_boston	10,571 (9)
		(22)02_new_york	11,589 (10)
		(23)03_philadelphia	6,507 (5)
		(24)04_atlanta	11,972 (10)
		(25)05_chicago	43,347 (36)
		(26)06_dallas	15,734 (13)
		(27)07_kansas_city	5,064 (4)
		(28)08_denver	9,832 (8)
		(29)09_san_francisco	2,411 (2)
		(30)10_seattle	1,820 (2)
Year	Year of sampling	Continuous (integer) 1984 to 2009	
SIC code	Constructed from the aggregation of the 4-digit SIC category	Nominal (median of 35 categories; IQR=[15;99])[b]	

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

[b] Median number of categories for the variable across chemicals and interquartile range

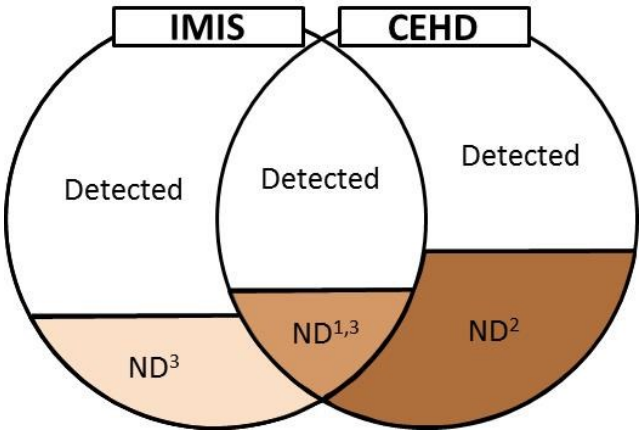


Figure 1: Partial overlap in records between IMIS and CEHD datasets

1. Non-detected dataset used for development of CART models
2. Non-detected dataset used for evaluation of external performance of CART models
3. Non-detected dataset predicted by CART models

Table II: Example of a confusion table for chemical agent

	ST Predicted	TWA Predicted	Model Error
ST	75	25	0.25
TWA	35	150	0.19
Use Error	0.32	0.14	Overall Error = 0.21

Table III: Number of samples (and proportion of TWA) for each chemical agent, stratified by dataset

Chemical agent	'Overlapping' dataset		'CEHD_only' dataset	
	Detected	ND (used for CART model building)	Detected	ND (used for CART external validation)
Group 1[a]				
Manganese fume	13028 (31)	3212 (88)	10872 (97)	8641 (94)
Vanadium Fume	1835 (23)	8638 (91)	1602 (97)	21252 (96)
Beryllium	813 (75)	9466 (91)	574 (98)	22725 (96)
Xylene	4761 (71)	1124 (35)	3197 (83)	715 (75)
2-Butanone	4887 (84)	876 (34)	1409 (89)	309 (82)
Cadmium fume	973 (66)	4729 (91)	1008 (98)	9559 (96)
Asbestos (all forms)	1957 (67)	2777 (60)	1331 (86)	2385 (78)
Formaldehyde	2237 (58)	1100 (35)	1495 (77)	776 (53)
Methylene Chloride	2485 (41)	531 (29)	983 (74)	205 (67)
Acetone	1829 (70)	467 (26)	1011 (84)	210 (76)
Isopropyl Alcohol	1577 (70)	477 (24)	965 (83)	228 (73)
Chromic acid	908 (44)	854 (58)	716 (83)	792 (79)
Hexone	1187 (67)	465 (36)	773 (89)	280 (80)
N-Butyl Acetate	1220 (75)	384 (32)	816 (85)	241 (80)
Benzene	388 (47)	951 (36)	297 (84)	610 (80)
Methyl Chloroform	1074 (60)	208 (20)	657 (80)	111 (59)
Ethyl benzene	899 (79)	364 (42)	616 (88)	222 (78)

2,6'-TDI	234 (39)	1029 (10)	417 (46)	726 (31)
HDI	274 (54)	834 (11)	305 (47)	664 (32)
Vinyl chloride	182 (60)	716 (21)	29 (72)	136 (88)
Sodium hydroxide	410 (62)	189 (61)	327 (83)	154 (76)
N-Butyl alcohol	385 (58)	201 (42)	630 (89)	369 (83)
Ammonia	396 (60)	168 (34)	329 (84)	147 (64)
Ethylene oxide	299 (70)	177 (39)	256 (77)	149 (66)
Methyl alcohol	252 (80)	211 (15)	184 (84)	71 (72)
Cadmium dust	228 (83)	165 (76)	161 (98)	139 (86)
Sulfur dioxide	294 (69)	85 (66)	204 (93)	83 (82)
Nitric acid	164 (84)	209 (64)	79 (87)	202 (76)
Mercury	287 (84)	74 (72)	330 (96)	49 (80)
Group 2[b]				
Lead, inorganic	17750 (99)	14867 (91)	4581 (97)	12632 (95)
Copper fume	11030 (99)	3606 (88)	11130 (98)	8610 (93)
Zinc oxide fume	10313 (97)	3319 (90)	11400 (97)	8806 (94)
Chromium	6429 (98)	5814 (91)	6898 (97)	14507 (95)
Nickel	4028 (99)	7336 (91)	4381 (98)	17499 (95)
Cobalt	1768 (99)	8746 (91)	1823 (97)	20721 (96)
Antimony	743 (99)	9455 (91)	868 (98)	22745 (96)
Molybdenum	1442 (99)	8410 (91)	1854 (98)	20778 (96)
Silica, quartz	6969 (99)	2384 (89)	4614 (95)	4629 (91)
4,4'-MDI	1378 (4)	2376 (7)	427 (24)	953 (32)
Stoddard Solvent	1114 (97)	1046 (28)	803 (85)	471 (77)

Petroleum naphtha	780 (92)	1118 (27)	537 (87)	450 (83)
2,4'-TDI	643 (8)	1158 (10)	291 (45)	826 (34)
Arsenic	673 (99)	633 (92)	1602 (99)	1631 (97)
Hydrogen chloride	555 (8)	509 (16)	177 (47)	225 (35)
2-Butoxyethanol	476 (99)	291 (46)	385 (91)	211 (82)
Coal tar pitch vol	499 (95)	221 (86)	461 (94)	144 (86)
Silica, cristobalite	48 (100)	633 (95)	24 (100)	1643 (96)
Hexane	410 (92)	211 (17)	377 (89)	86 (76)
Sulfuric acid	265 (100)	298 (81)	279 (93)	249 (84)
Ethyl alcohol	251 (98)	232 (16)	396 (90)	161 (77)
Phenol	266 (92)	212 (75)	170 (95)	131 (88)
Silver	334 (99)	105 (83)	412 (98)	626 (97)
Ethyl acetate	223 (100)	83 (22)	599 (89)	188 (82)
Fluoride	85 (100)	136 (54)	131 (91)	149 (72)
Copper dusts	154 (95)	32 (75)	568 (100)	121 (93)
Tin	63 (98)	103 (93)	301 (99)	1207 (96)
Group 3[c]				
Iron oxide fume	14193 (98)	1503 (87)	13802 (97)	3580 (92)
Toluene	8169 (42)	1025 (34)	3896 (81)	458 (73)
Styrene	4473 (34)	385 (30)	1626 (82)	207 (73)
PNOR (total dust)	2671 (99)	92 (84)	6246 (98)	3418 (81)
Tetrachloroethylene	1282 (37)	176 (19)	571 (70)	107 (71)
PNOR (resp dust)	1094 (100)	56 (82)	3039 (97)	18889 (96)
Trichloroethylene	700 (30)	47 (23)	374 (76)	64 (62)

[a]More or equal than 15% of NDs in 'overlapping' dataset, and between 10% and 90% of TWA in 'overlapping' detected dataset
[b]More or equal than 15% of NDs in 'overlapping'dataset, and more than 90% of TWA or ST in 'overlapping' detected dataset
[c]Less than 15% of NDs in 'overlapping'dataset

Table IV: Importance of variables in CART models across the 54 chemical agents

	Most important in model (n)[a]	Present in model (n)[b]	Relative Importance (%) (median and [IQR])[c]
Inspection type	0	54	7 [5;9]
Inspection scope	0	47	3 [2;4]
Union status	0	46	3 [1;4]
OSHA plan	0	20	0 [0;1]
OSHA region	1	54	16 [14;22]
Year	4	53	14 [12;19]
SIC code	49	54	52 [44;59]

[a]Number of CART models in which variable was the most important
[b]Number of CART models in which variable was present
[c]Median relative importance of variable across CART models and interquartile range

Table V: Classification performance of CART model and two simpler assignment methods for each agent

Chemical agent	CART model		Simpler methods	
	Bagging error[a]	**CEHD prediction error**	**Random assignment[b] error**	**Assignment of most frequent category[c] error**
Group 1[d]				
Manganese fume	0,156	0,118	0,656	0,877
Vanadium Fume	0,104	0,08	0,724	0,909
Beryllium	0,102	0,076	0,297	0,092
Xylene	0,254	0,447	0,578	0,655
2-Butanone	0,199	0,433	0,619	0,656
Cadmium fume	0,101	0,062	0,367	0,091
Asbestos (all forms)	0,294	0,336	0,477	0,4
Formaldehyde	0,289	0,473	0,545	0,648
Methylene Chloride	0,224	0,532	0,429	0,286
Acetone	0,225	0,605	0,597	0,739
Isopropyl Alcohol	0,216	0,43	0,585	0,761
Chromic acid	0,22	0,323	0,496	0,584
Hexone	0,239	0,465	0,578	0,641
N-Butyl Acetate	0,274	0,544	0,615	0,682
Benzene	0,283	0,437	0,526	0,355
Methyl Chloroform	0,231	0,376	0,558	0,803
Ethyl benzene	0,305	0,597	0,53	0,58
2,6'-TDI	0,113	0,329	0,431	0,102

HDI	0,111	0,345	0,52	0,892
Vinyl chloride	0,111	0,695	0,557	0,789
Sodium hydroxide	0,313	0,315	0,418	0,386
N-Butyl alcohol	0,306	0,475	0,542	0,577
Ammonia	0,253	0,396	0,53	0,661
Ethylene oxide	0,257	0,481	0,542	0,61
Methyl alcohol	0,106	0,618	0,697	0,853
Cadmium dust	0,119	0,209	0,352	0,242
Nitric acid	0,325	0,362	0,411	0,359
Group 2[e]				
Lead, inorganic	0,098	0,087	0,095	0,088
Copper fume	0,14	0,113	0,124	0,116
Zinc oxide fume	0,123	0,11	0,124	0,102
Chromium	0,107	0,09	0,111	0,092
Nickel	0,094	0,068	0,093	0,086
Cobalt	0,098	0,074	0,098	0,088
Antimony	0,101	0,067	0,098	0,09
Molybdenum	0,103	0,085	0,099	0,09
Silica, quartz	0,099	0,115	0,123	0,11
4,4'-MDI	0,082	0,322	0,1	0,066
Stoddard Solvent	0,249	0,559	0,713	0,723
Petroleum naphtha	0,214	0,522	0,695	0,729
2,4'-TDI	0,133	0,37	0,17	0,104
Arsenic	0,088	0,043	0,082	0,076

Hydrogen chloride	0,139	0,382	0,214	0,161
2-Butoxyethanol	0,286	0,507	0,529	0,536
Coal tar pitch vol	0,151	0,195	0,176	0,14
Silica, cristobalite	0,064	0,045	0,049	0,049
Hexane	0,1	0,325	0,82	0,829
Sulfuric acid	0,195	0,242	0,195	0,195
Ethyl alcohol	0,147	0,707	0,832	0,841
Phenol	0,229	0,331	0,307	0,255
Fluoride	0,339	0,278	0,456	0,456
Group 3[f]				
Iron oxide fume	0,156	0,157	0,15	0,128
Toluene	0,28	0,52	0,461	0,341
Styrene	0,223	0,498	0,431	0,304
Tetrachloroethylene	0,192	0,694	0,386	0,193

[a]Error: proportion of ND results incorrectly classified by the model
[b]Random assignement of TWA or ST based on probability from detected dataset
[c]Assignement of most frequent category (TWA, ST) calculated from detected dataset
[d]More or equal than 15% of NDs in 'overlapping' dataset, and between 10% and 90% of TWA in 'overlapping' detected dataset
[e]More or equal than 15% of NDs in 'overlapping'dataset, and more than 90% of TWA or ST in 'overlapping' detected dataset
[f]Less than 15% of NDs in 'overlapping'dataset

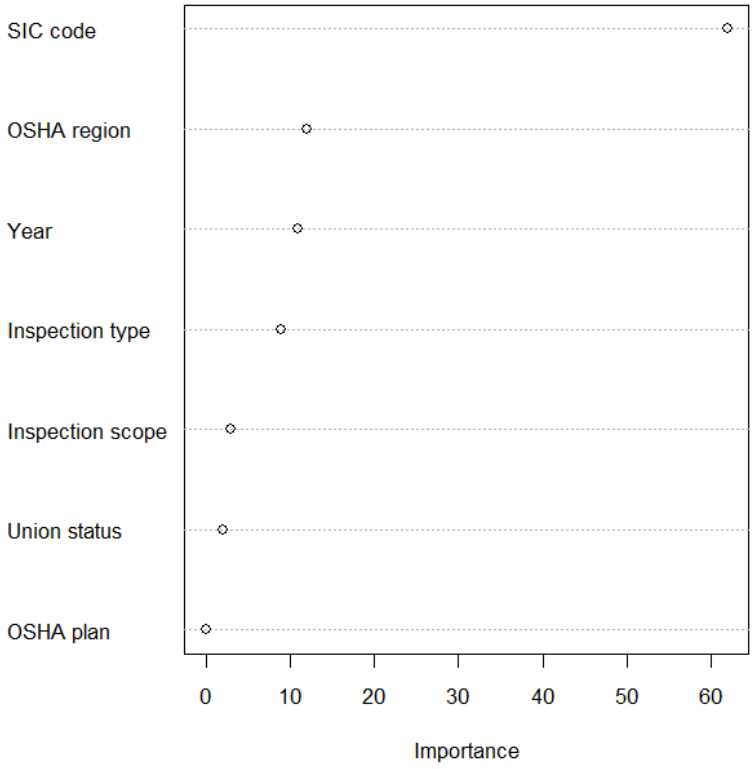


Figure 2: Plot for variable importance in CART model for manganese fume

Table VI: Number of sample results and proportion of TWA in IMIS for each chemical agent

Chemical agent	IMIS detected	IMIS ND [predicted][a]
Group 1[b]		
Manganese fume	19004 (50)	4601 (92)
Vanadium Fume	1481 (40)	9986 (96)
Beryllium	2004 (50)	11629 (97)
Xylene	10811 (86)	3139 (54)
2-Butanone	5016 (86)	1748 (55)
Cadmium fume	2770 (71)	7192 (97)
Asbestos (all forms)	5426 (76)	4948 (69)
Formaldehyde	6603 (66)	2267 (35)
Methylene Chloride	4149 (58)	727 (31)
Acetone	4716 (85)	1198 (50)
Isopropyl Alcohol	3084 (86)	832 (68)
Chromic acid	2088 (70)	1674 (71)
Hexone	2363 (85)	889 (54)
N-Butyl Acetate	2980 (86)	832 (54)
Benzene	1169 (69)	1775 (57)
Methyl Chloroform	2292 (83)	424 (29)
Ethyl benzene	2305 (86)	883 (56)
2,6'-TDI	172 (67)	551 (6)
HDI	654 (58)	844 (11)
Vinyl chloride	206 (80)	428 (43)

Sodium hydroxide	830 (79)	531 (74)
N-Butyl alcohol	434 (76)	283 (58)
Ammonia	868 (79)	296 (33)
Ethylene oxide	553 (73)	276 (28)
Methyl alcohol	485 (82)	207 (20)
Cadmium dust	1578 (92)	1615 (95)
Nitric acid	376 (87)	493 (64)
Group 2[c]		
Lead, inorganic	31519 (99)	25401 (97)
Copper fume	19062 (99)	5680 (96)
Zinc oxide fume	19135 (87)	4829 (95)
Chromium	10330 (98)	9040 (96)
Nickel	6790 (98)	10816 (98)
Cobalt	3270 (98)	11059 (96)
Antimony	1177 (98)	11787 (97)
Molybdenum	1839 (98)	9914 (96)
Silica, quartz	20284 (99)	5969 (95)
4,4'-MDI	1650 (15)	2571 (3)
Stoddard Solvent	2451 (98)	991 (43)
Petroleum naphtha	3089 (95)	1542 (51)
2,4'-TDI	758 (28)	1185 (8)
Arsenic	1607 (99)	1139 (99)
Hydrogen chloride	693 (24)	727 (12)
2-Butoxyethanol	925 (98)	576 (60)

Coal tar pitch vol	1862 (98)	669 (88)
Silica, cristobalite	357 (97)	1846 (99)
Hexane	1348 (97)	406 (48)
Sulfuric acid	915 (98)	897 (80)
Ethyl alcohol	466 (94)	232 (21)
Phenol	618 (98)	425 (91)
Fluoride	573 (100)	321 (89)
Group 3[d]		
Iron oxide fume	28471 (99)	2488 (94)
Toluene	18386 (63)	3166 (45)
Styrene	8150 (58)	829 (61)
Tetrachloroethylene	2326 (59)	305 (36)

[a]Predicted by CART model
[b]More or equal than 15% of NDs in 'overlapping' dataset, and between 10% and 90% of TWA in 'overlapping' detected dataset
[c]More or equal than 15% of NDs in 'overlapping'dataset, and more than 90% of TWA or ST in 'overlapping' detected dataset
[d]Less than 15% of NDs in 'overlapping'dataset

Table VII: Comparison of odds ratios (ORs) of a sample result exceeding the TLV ± exposure type variable for 3 chemical agents

Variable/ category	2-Butanone		Formaldehyde		Xylene	
	Without expo_type OR (95% CI)	With expo_type OR (95% CI)	Without expo_type OR (95% CI)	With expo_type OR (95% CI)	Without expo_type OR (95% CI)	With expo_type OR (95% CI)
Inspection type						
Planned	1.00 (reference)[a]	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)
Complaint	0.99 (0.68; 1.45)[b]	0.96 (0.65; 1.43)	1.28 (1.08;1.53)	1.32 (1.11;1.58)	1.12 (0.73;1.74)	1.11 (0.71; 1.73)
Follow-up	2.92 (1.56; 5.48)	2.56 (1.33; 4.91)	1.65 (1.14;2.39)	1.74 (1.20;2.52)	2.16 (0.88;5.33)	2.12 (0.85; 5.30)
Referral	1.57 (0.98; 2.51)	1.32 (0.80; 2.15)	1.61 (1.29;2.01)	1.64 (1.31;2.04)	1.95 (1.20;3.17)	1.76 (1.07; 2.89)
OSHA plan						
Federal	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)
State	0.54 (0.37; 0.80)	0.53 (0.35; 0.80)	0.48 (0.42;0.56)	0.51 (0.44;0.59)	0.62 (0.42;0.91)	0.60 (0.40; 0.89)
Inspection scope						
Comprehensive	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)
Partial	0.94 (0.69; 1.28)	0.95 (0.69; 1.31)	0.86 (0.74;1.01)	0.86 (0.73;1.00)	0.84 (0.59;1.19)	0.88 (0.62; 1.26)
Union status						
No	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)
Yes	0.79 (0.57; 1.10)	0.79 (0.57; 1.10)	1.11 (0.98;1.27)	1.07 (0.94;1.23)	0.72 (0.48;1.09)	0.71 (0.47; 1.07)
OSHA region[c]						
01_boston	1.76 (1.02; 3.05)	1.64 (0.94; 2.87)	1.07 (0.83;1.38)	1.20 (0.92;1.56)	1.34 (0.70;2.56)	1.45 (0.76; 2.78)
02_new_york	2.48 (1.42; 4.34)	2.62 (1.49; 4.60)	1.43 (1.17;1.75)	1.50 (1.22;1.84)	0.79 (0.37;1.70)	0.83 (0.39; 1.78)
03_philadelphia	2.13 (1.13; 3.99)	2.12 (1.13; 4.01)	1.39 (1.08;1.78)	1.42 (1.10;1.83)	2.69 (1.50;4.82)	2.89 (1.61; 5.21)
04_atlanta	2.55 (1.56; 4.16)	2.49 (1.52; 4.09)	1.58 (1.34;1.87)	1.60 (1.35;1.90)	1.59 (0.97;2.59)	1.63 (0.99; 2.66)
05_chicago	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)
06_dallas	1.92 (1.09; 3.38)	1.95 (1.09; 3.47)	0.93 (0.74;1.18)	0.94 (0.74;1.19)	0.70 (0.34;1.43)	0.70 (0.34; 1.43)
07_kansas_city	1.74 (0.94; 3.25)	1.76 (0.94; 3.31)	1.56 (1.15;2.12)	1.71 (1.25;2.33)	1.34 (0.67;2.70)	1.45 (0.72; 2.92)

08_denver	1.79 (0.79; 4.08)	1.76 (0.77; 4.02)	0.71 (0.51;1.00)	0.66 (0.47;0.93)	1.53 (0.72;3.23)	1.53 (0.72; 3.27)
09_san_francisco	1.28 (0.37; 4.44)	1.48 (0.42; 5.18)	2.93 (2.07;4.15)	3.04 (2.13;4.32)	3.87 (1.51;9.90)	4.90 (1.90;12.66)
10_seattle	8.85 (4.67;16.75)	8.22 (4.21;16.05)	2.30 (1.75;3.03)	2.38 (1.80;3.15)	5.16 (2.96;9.01)	4.74 (2.67; 8.43)
Establishment size						
Small (1-35)	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)
Medium (36-150)	0.69 (0.47; 1.01)	0.70 (0.47; 1.03)	0.84 (0.72;0.98)	0.85 (0.73;1.00)	0.74 (0.51;1.06)	0.77 (0.53; 1.12)
Large (151+)	1.38 (0.93; 2.05)	1.37 (0.92; 2.04)	0.74 (0.62;0.88)	0.75 (0.63;0.89)	0.42 (0.25;0.71)	0.42 (0.25; 0.71)
Penalty						
None	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)	1.00 (reference)
Low	3.36 (1.77; 6.38)	3.25 (1.70; 6.19)	1.06 (0.89;1.25)	1.07 (0.90;1.27)	1.46 (0.83;2.57)	1.39 (0.79; 2.46)
Medium	4.31 (2.20; 8.45)	3.90 (1.98; 7.67)	1.25 (1.04;1.51)	1.26 (1.05;1.52)	2.03 (1.12;3.68)	1.88 (1.04; 3.41)
High	5.64 (2.81;11.35)	5.23 (2.59;10.59)	1.39 (1.13;1.71)	1.40 (1.14;1.73)	2.35 (1.22;4.50)	2.29 (1.19; 4.41)
Exposure type						
Short term	-	1.00 (reference)	-	1.00 (reference)	-	1.00 (reference)
TWA	-	0.51 (0.35; 0.75)	-	0.59 (0.53;0.67)	-	0.33 (0.22; 0.49)

[a]OR of the reference levels taken as 1.
[b]95% confidence interval.
[c]https://www.osha.gov/html/RAmap.html.

5.7 References

Berk R. (2008) Statistical Learning from a Regression Perspective. Springer Series in Statistics. 978-0-387-77500-5.

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

Burstyn I, Slutsky A, Lee DG *et al.* (2014) Beyond crosswalks: reliability of exposure assessment following automated coding of free-text job descriptions for occupational epidemiology. Ann Occup Hyg; 58: 482-92.

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.

Friesen MC, Pronk A, Wheeler DC *et al.* (2013) Comparison of algorithm-based estimates of occupational diesel exhaust exposure to those of multiple independent raters in a population-based case-control study. Ann Occup Hyg; 57: 470-81.

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, Wegman DH, Dellenbaugh CA. (1986) An approach to the characterization of silica exposure in U.S. industry. Am J Ind Med; 10: 345-61.

Gomez 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. 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.

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.

Lavoue J, Gerin 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.

Leblanc M, Crowley J. (1993) Survival Trees by Goodness of Split. Journal of the American Statistical Association; 88: 457-67.

Lee D, Lavoue J, Spinelli J *et al.* (2015) Statistical modeling of occupational exposure to polycyclic aromatic hydrocarbons using OSHA data. J Occup Environ Hyg; 1-14.

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: 378-83.

Melville R, Lippmann M. (2001) Influence of data elements in OSHA air sampling database on occupational exposure levels. Appl Occup Environ Hyg; 16: 884–99.

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

OSHA. (2015) Chemical Exposure Health Data. https://www.osha.gov/opengov/healthsamples.html (Accessed 2 August 2015).

Raileanu LE, Stoffel K. (2004) Theoretical Comparison between the Gini Index and Information Gain Criteria. Annals of Mathematics and Artificial Intelligence; 41: 77-93.

Rajan B, Alesbury R, Carton B *et al.* (1997) European proposal for core information for the storage and exchange of workplace exposure measurements on chemical agents. Appl Occup Environ Hyg; 12: 31-9.

Ripley B. (2015) Recursive Partitioning and Regression Trees. https://cran.r-project.org/web/packages/rpart/rpart.pdf (Accessed 20 october 2014).

Russ DE, Ho KY, Johnson CA *et al.* (2014) Computer-Based Coding of Occupation Codes for Epidemiological Analyses. Proc IEEE Int Symp Comput Based Med Syst; 2014: 347-50.

Sarazin P, Burstyn I, Kincl L *et al.* (2016a) Characterization of the selective recording of sample results in OSHA's IMIS databank, 1984-2009: statistical modeling of ancillary information across 78 chemicals. Manuscript in preparation.

Sarazin P, Burstyn I, Kincl L *et al.* (2016b) Trends in OSHA Compliance Monitoring Data 1979-2011: Statistical Modeling of Ancillary Information across 77 Chemicals. Ann Occup Hyg.

Speiser JL, Durkalski VL, Lee WM. (2014) Random forest classification of etiologies for an orphan disease. Stat Med.

Stewart PA, Rice C. (1990) A source of exposure data for occupational epidemiology studies. Appl Occup Environ Hyg; 5: 359-63.

Strobl C, Malley J, Tutz G. (2009) An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol Methods; 14: 323-48.

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

Van Hulst A, Roy-Gagnon MH, Gauvin L *et al.* (2015) Identifying risk profiles for childhood obesity using recursive partitioning based on individual, familial, and neighborhood environment factors. Int J Behav Nutr Phys Act; 12: 17.

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

Wheeler DC, Burstyn I, Vermeulen R *et al.* (2013) Inside the black box: starting to uncover the underlying decision rules used in a one-by-one expert assessment of occupational exposure in case-control studies. Occup Environ Med; 70: 203-10.

Wood S. (2014) Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation. Available at http://cran.r-project.org/web/packages/mgcv/index.html (Accessed 15 september 2014).

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.

CHAPITRE 6- Discussion générale

Discussion générale

Cette recherche visait à documenter l'existence de biais potentiels reliés à la sélection des secteurs d'activité et des établissements où des mesures sont effectuées, aux stratégies de mesure employées et à l'enregistrement des résultats d'exposition dans la banque IMIS. De plus, une procédure transparente et reproductible a été proposée facilitant l'interprétation de la proportion élevée de résultats ND dans la banque. La figure 1 présente les sections du schéma conceptuel des biais dans la banque IMIS visées par les chapitres 3 et 4 de la thèse, alors que le tableau I présente un résumé succinct des résultats classés par chapitre. Les sections suivantes discutent des résultats et contributions de la thèse, présentent les limites et forces globales de l'étude, les implications potentielles pour la santé publique et les perspectives futures de recherche.

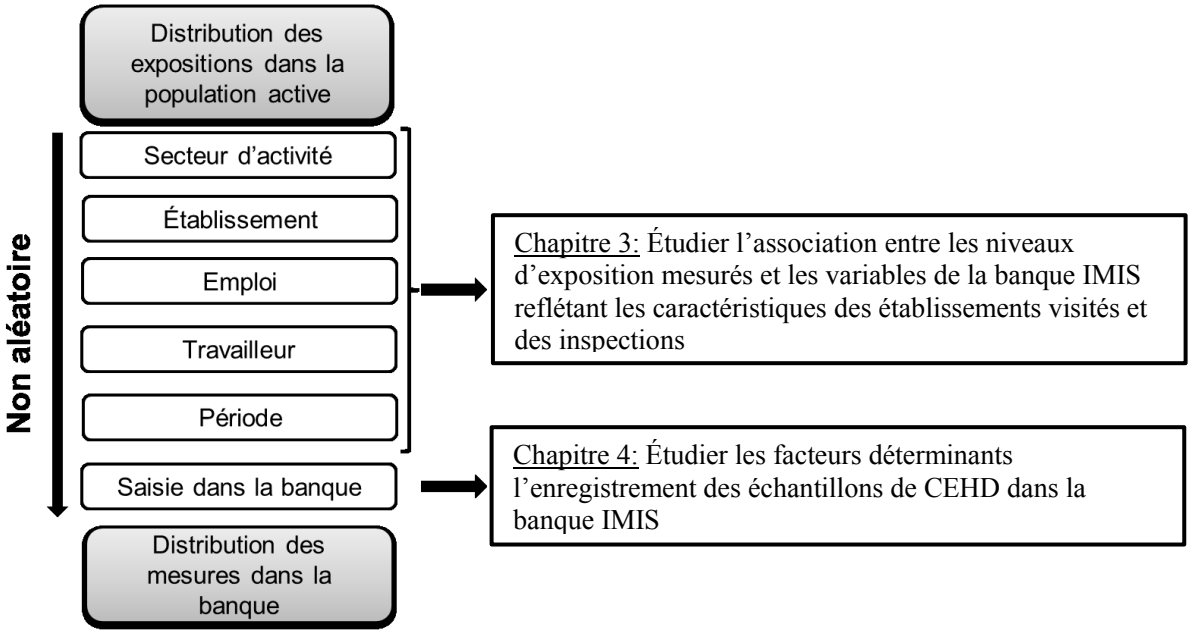


Figure 1 : Sections du schéma conceptuel des biais dans la banque IMIS visées par chaque chapitre de la thèse

Table I : Synthèse des principaux résultats obtenus lors de l'analyse de la banque IMIS – étude des biais (chapitres 3 et 4) et interprétation des mesures non détectées (chapitre 5)

Chapitre 3: Trends in OSHA compliance monitoring data 1979-2011: statistical modeling of ancillary information across 77 chemicals

Méthodes d'analyse	Résultats principaux[a]
• Banque de données IMIS • 77 agents chimiques • 10 variables analysées • Modèles GAM – régression logistique de la probabilité de dépasser la TLV - 511 047 mesures personnelles • Modèles GAM – modèles linéaires mixtes pour les mesures détectées - 299 791 mesures personnelles • Inférence multimodèle • Méta-analyse pour la synthèse des résultats à travers les agents • Période 1979 – 2011	• Données visites de suivi > données visites planifiées (facteur ~1,6 à travers les agents) • Données établissements avec historique de non-conformité > données établissements conformes (facteur ~1,5 à travers les agents) • Données régime fédéral > données régime étatique (facteur ~1,25 à travers les agents) • Données plaintes ~ données visites planifiées • Données courte durée > données VEMP-8h (facteur ~3 à travers les agents)[b] • Variable région influente • Tendance historique : réduction de l'exposition pour 50/77 agents de 1979 à 2011

Chapitre 4: Characterization of the selective recording of sample results in OSHA's IMIS databank, 1984-2009: statistical modeling of ancillary information across 78 chemicals

Méthodes d'analyse	Résultats principaux
• Croisement des banques de données IMIS et CEHD • 78 agents chimiques • 9 variables analysées • Régression de Poisson modifiée de la probabilité d'enregistrement dans IMIS - 588 818 mesures personnelles • Période 1984 – 2009	• 38% des résultats CEHD enregistrés dans IMIS • Enregistrement résultats détectés > non détectés (facteur ~1,7 pour agents mesurés sur un panel; facteur ~1,2 pour agents mesurés seuls) • Tendance historique : augmentation de l'enregistrement de 1984 à 2009 pour agents mesurés seuls; constant pour agents mesurés sur un panel • Variation de l'enregistrement selon l'agent chimique et la région • Enregistrement n'est pas associé à l'émission d'une infraction pour le dépassement de la norme

Chapitre 5: Non-detects in OSHA's IMIS databank, are they short term or 8-hour shift-long samples? Prediction for 54 chemicals using recursive partitioning statistical methods

Méthodes d'analyse	Résultats principaux
• Croisement des banques de données IMIS et CEHD • 54 agents chimiques • Modèles CART pour prédire le type d'exposition (courte durée, VEMP-8h) des mesures non détectées dans IMIS • 7 variables utilisées pour la prédiction • Période 1984 – 2009	• Tableau de prédiction du type d'exposition (courte durée, VEMP-8h) pour 54 agents • Code SIC, région et année sont les variables les plus importantes pour prédire le type d'exposition • Meilleure performance des modèles CART comparée à attribution aléatoire (pour tous les agents (27/27) ayant proportion substantielle de mesures ND, courte durée et VEMP-8h) • Meilleure performance des modèles CART comparée à attribution de la catégorie la plus fréquente (pour 23 des 27 agents ayant proportion substantielle de mesures ND, courte durée et VEMP-8h)

[a]Régression logistique; [b]Modèles linéaires

Dans la première analyse présentée au chapitre 3, nous avons examiné l'association entre les niveaux d'exposition mesurés et les variables de la banque IMIS reflétant les caractéristiques des établissements visités et des inspections. L'approche de méta-analyse a été utilisée afin de synthétiser l'effet de chaque variable à travers l'ensemble des agents chimiques. Globalement, l'étude a montré l'existence d'associations entre les niveaux d'exposition et plusieurs variables internes IMIS.

Dans la deuxième analyse, les mécanismes par lesquels les résultats d'échantillonnage sont enregistrés dans la banque IMIS ont été examinés en effectuant le croisement avec la banque de résultats d'analyse CEHD pour l'ensemble des agents chimiques communs aux deux banques. Globalement, les résultats de l'étude ont montré que le sous-enregistrement des résultats dans IMIS est important (moyenne 38% transmis), et qu'il dépend du niveau d'exposition: les résultats ND ont une probabilité plus faible d'enregistrement dans IMIS que les résultats détectés.

Finalement, l'approche CART a été utilisée dans la troisième analyse afin de prédire, parmi les résultats non détectés (ND) de la banque IMIS, lesquels correspondent à des mesures courte durée ou VEMP-8h, facilitant leur analyse future. L'étude a montré que les modèles CART fournissent un moyen systématique et plus performant que les approches existantes permettant d'effectuer les prédictions, particulièrement pour les agents chimiques pour lesquels ce type de prédiction est essentiel (proportion élevée de résultats ND, et partage équilibré entre mesures courte durée et mesures VEMP-8h).

6.1 Contributions de la recherche

6.1.1 Représentativité de la banque IMIS

6.1.1.1 Association des niveaux d'exposition avec des variables internes

Ne disposant pas de source de référence (gold standard) à laquelle comparer les niveaux d'exposition présents dans la banque IMIS, nos analyses ne peuvent fournir directement d'information sur la représentativité des niveaux d'exposition dans IMIS. L'étude de la différence potentielle entre les expositions rencontrées dans IMIS et l'ensemble des milieux de travail américains a donc été réalisée par l'étude de l'association entre les niveaux d'exposition enregistrés dans la banque et des éléments contextuels potentiellement associés à la stratégie de mesure.

Deux grandes catégories de variables ont été analysées : celles reflétant les caractéristiques des établissements visités (p. ex. taille de l'établissement, présence d'un syndicat) et celles reflétant les caractéristiques de l'inspection (p. ex. raison de la visite, régime OSHA – fédéral/d'État). Nos analyses ont permis d'identifier des associations entre les niveaux d'exposition et certaines variables (voir ci-dessous) pour chacune de ces catégories. Ces associations étaient homogènes à travers les agents chimiques, ce qui suggère l'existence de tendances sous-jacentes globales dans IMIS.

Les niveaux d'exposition étaient généralement plus élevés lors des visites de suivi effectués par les inspecteurs comparé aux visites planifiées. Ce résultat reflète le biais de sélection où les inspecteurs sont davantage susceptibles de retourner prendre des mesures dans les milieux de travail où des problèmes reliés à l'exposition avaient préalablement été identifiés. En revanche, notre étude suggère l'absence d'un biais vers le haut des expositions résultant de

visites causées par des plaintes d'employés comparé aux visites planifiées, une hypothèse soulevée à maintes reprises dans la littérature depuis l'implantation de la banque IMIS (Froines et coll., 1986; Froines et coll., 1990; Gomez, 1997; Melville et Lippmann, 2001; Lavoue et coll., 2008; Henn et coll., 2011; Lavoue et coll., 2013). Le total des amendes reçues par un établissement étaient également associé aux niveaux d'exposition dans notre étude. De plus, cette association était indépendante de la nature des infractions, ce qui suggère que les comportements non-conformes (même si non-reliés à l'exposition chimique) étaient prédicteurs de niveaux d'exposition élevés. Finalement, les niveaux d'exposition étaient généralement plus faibles lors des inspections menées sous un régime OSHA d'État en comparaison au régime OSHA fédéral. Cette tendance pourrait s'expliquer par des priorités ou des pratiques différentes entre les deux stades de gouvernement.

Les résultats de nos modèles ont montré que l'ampleur des associations entre les niveaux d'exposition et les variables internes dans IMIS était du même ordre de grandeur que celles retrouvées dans des études basées sur d'autres jeux de données d'exposition. Par exemple, Lavoue et al. (2005) ont montré que les niveaux d'exposition au formaldéhyde étaient en moyenne 3 fois plus élevés pour des données gouvernementales comparées à des mesures prises par une équipe de recherche dans l'industrie des panneaux de bois aggloméré au Québec. Peters et coll. (2011) ont montré que les niveaux d'exposition à la silice cristalline étaient 1,77 fois plus élevé pour la stratégie de mesure « pire des cas » comparé à une stratégie de mesure représentative dans la banque de données ExpoSYN. Notre étude a montré que les niveaux d'exposition étaient en moyenne 1,06 plus élevés dans les inspections de suivi comparé aux inspections planifiées, alors qu'ils étaient en moyenne 1,16 fois plus élevés lors des mesures effectuées sous un régime OSHA fédéral par rapport au régime OSHA d'État. Les niveaux d'exposition augmentaient également avec le nombre total des

amendes reçues par un établissement (méta-RIE de 1,18 entre les catégories « élevée » et « aucune »). Les associations observées dans notre étude semblent en général de plus faible amplitude, mais ceci s'explique par l'emploi de la méta-analyse qui rapporte un effet moyen à travers un large ensemble d'agents. Par exemple, le méta-RIE de 1,18 correspondait à l'agrégation de RIE variant entre 0,94 et 2,71 à travers les agents.

En résumé, nos résultats ont montré la présence de plusieurs mécanismes de sélection dans le processus conduisant à l'enregistrement d'une mesure d'exposition dans IMIS, ce qui suggère l'existence de différences systématiques entre les niveaux rapportés dans les banques OSHA et les niveaux moyens d'exposition dans la population de travailleurs. Les deux exemples qui suivent permettent d'illustrer de quelle façon nos analyses permettent de prendre en considération ces biais afin d'améliorer les estimés d'exposition effectués à partir d'IMIS.

1) Pour les variables reflétant les caractéristiques des établissements visités, il est possible de pondérer les estimés d'exposition effectués à partir d'IMIS en fonction de la distribution de la variable dans la population. Par exemple, pour la variable taille de l'établissement, les estimés d'exposition devraient être pondérés en fonction de la distribution des tailles d'entreprises dans la population.

2) Les autres variables reflètent quant à elles des stratégies de sélection et nécessitent donc de porter un jugement sur laquelle des stratégies est la plus représentative de l'ensemble des milieux de travail. À titre d'exemple, les niveaux d'exposition des échantillons recueillis sous un régime OSHA fédéral correspondent-ils davantage aux niveaux dans la population que ceux recueillis sous un régime OSHA d'État? Pour cette variable, il s'avère nécessaire de

porter un jugement sur laquelle des deux situations est la plus représentative de la population (ou de simplement considérer les deux situations comme étant équivalentes). Concrètement, l'estimation des niveaux d'exposition pour un agent chimique en particulier peut être réalisée au choix de l'utilisateur. Si l'on considère équivalentes les catégories « régime OSHA fédéral » et « régime OSHA d'État », le scénario de prédiction pourrait être élaboré avec une pondération équilibrée (50% OSHA fédéral et 50% OSHA d'État).

6.1.1.2 Enregistrement des résultats dans la banque IMIS

Des études antérieures avaient suggéré l'existence dans la banque IMIS d'un sous-enregistrement des résultats mesurés par les inspecteurs d'OSHA (Mendeloff, 1984; Jones et coll., 1986; Lavoue et coll., 2013), qui correspond à la case « Saisie dans la banque » de la figure 1. Il était mentionné que la décision d'enregistrer un résultat dans la banque IMIS pouvait de plus dépendre du niveau d'exposition, de l'agent chimique ou de l'émission d'une infraction liée à la surexposition. Ces études rapportaient cependant des résultats contradictoires sur l'existence d'un sous-enregistrement préférentiel des résultats faibles ou non détectés.

À partir du croisement entre la banque IMIS et la banque de résultats d'analyse CEHD pour l'ensemble des agents chimiques communs aux deux banques, seulement 38% des données CEHD peuvent être associées à une évaluation dans IMIS. Ce résultat est similaire à la proportion de résultats enregistrés de 50% rapportée par Mendeloff (1984) et Jones et coll. (1986) sur des données couvrant des périodes anciennes d'IMIS. Notre analyse supporte l'hypothèse selon laquelle le sous-enregistrement des résultats dans IMIS est différentiel : les résultats détectés ont une probabilité plus élevée d'enregistrement que les résultats ND

(facteurs de 1,7 et 1,2 pour les agents mesurés sur un panel et les agents mesurés seuls, respectivement). Le sous-enregistrement, en plus de limiter le nombre de valeurs utilisables pour établir des portraits d'exposition, pose un problème de représentativité des données IMIS par rapport à la population s'il s'avère différentiel. En effet, même si l'on considérait que les résultats des échantillons recueillis par les inspecteurs présents dans la banque d'analyse CEHD représentaient bien la population aux États-Unis, la sous-représentation des valeurs faibles (ND) dans IMIS causé par ce phénomène causerait un biais de surestimation. Nos résultats tendent de plus à démontrer que la probabilité d'enregistrement des résultats ND mesurés sur un panel d'agent (p. ex. panel de métaux) est encore plus faible que ceux mesurés seuls. Ceci suggère qu'une partie importante de ces mesures ont simplement été générées par la technique de laboratoire : chaque fois que l'inspecteur OSHA était intéressé à un agent chimique en particulier, des résultats ND étaient créés pour les autres agents du panel, et ces résultats tendaient à ne pas être transmis dans IMIS.

6.1.2 Approches d'analyse pour l'étude des banques IMIS et CEHD

6.1.2.1 Modélisation statistique des données d'exposition

L'étude des associations entre les niveaux d'exposition aux agents chimiques dans IMIS et les éléments contextuels associés à la stratégie de mesure ont été réalisées à partir de modèles statistiques de régression multiple (Burstyn et Teschke, 1999; Lavoue et coll., 2008; Hamm et Burstyn, 2011; Sauve et coll., 2013; Lee et coll., 2015). Cette méthode représentait un outil de choix pour l'analyse des données d'exposition puisqu'elle permet d'évaluer l'influence conjointe de multiples variables sur les niveaux d'exposition (chapitre 3).

L'analyse des données d'IMIS a été compliquée par la présence d'une proportion élevée de résultats non décelés. Compte tenu de la difficulté à interpréter un résultat sous la limite de détection comme étant « présent mais non détecté » ou « non présent », deux approches de modélisation ont été utilisées, chacune correspondant à l'une des interprétations précédentes (avec ou sans les ND). La modélisation de la probabilité que l'exposition dépasse la TLV (avec les ND – approche logistique) correspondait à l'interprétation « exposition présente mais non détecté », alors que la modélisation restreinte aux résultats détectés (approche linéaire) correspondait à l'interprétation « exposition non présente ». Cette approche parallèle est particulièrement d'intérêt dans l'étude des données IMIS puisqu'elle permet de vérifier si les résultats obtenus quant à l'effet d'une variable sont cohérents indépendamment des problèmes méthodologiques associés à chaque approche. Dans notre étude, les directions d'effet des variables étaient similaires quelle que soit l'approche de modélisation utilisée, ce qui renforce les conclusions de cette étude.

Il est toutefois complexe d'évaluer si les rapports de cotes (RC) et les indices relatifs d'exposition (RIE) rapportés pour les approches logistique et linéaire correspondaient à des ampleurs d'associations similaires avec les niveaux d'exposition. En effet, ces indices mesurent des quantités différentes. Il est tout-de-même possible d'obtenir une idée approximative si l'on suppose une distribution log-normale de référence pour les niveaux d'exposition. Cette approximation a été effectuée pour le plomb en prenant comme exemple la variable « total des pénalités » (« élevée » vs. « aucune »; RC = 1,96 et RIE = 1,67). Avec une proportion de dépassement brute de 21% pour la catégorie « aucune pénalité », la proportion de dépassement pour la catégorie « pénalité élevée » serait de 34% à partir du RC de l'approche logistique et de 29% à partir du RIE de l'approche linéaire. Il est à noter que ces estimés pourraient quelque peu varier selon le choix de la distribution de référence. Ils

permettent néanmoins de fournir un aperçu de la comparabilité des ampleurs d'associations rapportées par les deux approches de régression.

La régression de Poisson modifiée avec estimation robuste de la variance, une méthode similaire à la régression logistique, a été utilisée pour l'étude de l'enregistrement sélectif des résultats dans la banque IMIS (chapitre 4). Bien que cette méthode relativement récente n'ait été utilisée que très rarement dans l'étude des résultats d'exposition en hygiène du travail (Hamm et Burstyn, 2011; Zou et Donner, 2013), elle permet de quantifier l'effet de variables prédictives sur la variable réponse directement sous la forme d'un risque relatif plutôt qu'un rapport de cote. L'utilisation de cette approche est particulièrement indiquée et devrait être privilégiée lorsque la prévalence du caractère étudié est élevée dans la population, ce qui était le cas dans l'étude du chapitre 4 alors que la proportion de base d'enregistrement des données dans IMIS était considérée élevée (c.-à-d. >25%).

Il est à noter que l'emploi de la régression logistique pour modéliser la probabilité que les niveaux d'exposition dépassent la TLV dans le chapitre 3 a pu causer une surestimation des effets. En effet, lorsque la prévalence du caractère étudié est >25% (seuil approximatif), l'interprétation des rapports de cotes comme des risques relatifs est faussée car dans ce cas le RR sera plus faible que le RC (p. ex. un RC de 1.5 correspond à un RR de 1.3 lorsque la prévalence est de 30%). Dans notre analyse, la proportion des niveaux d'exposition dépassant la TLV était >25% pour moins de 20 agents sur 77, ce qui devrait avoir limité l'influence de cet effet sur les résultats méta-analytiques. Pour le chapitre 4, alors que la prévalence du caractère étudiée (c.-à-d. échantillon CEHD enregistré dans IMIS) était de 38%, la régression de Poisson modifiée a été sélectionnée pour la conduite des analyses.

6.1.2.2 L'approche de méta-analyse pour la synthèse des résultats à travers les agents

Un défi particulier rencontré pour cette recherche doctorale consistait à développer une approche permettant de combiner les résultats de tous les agents chimiques et d'obtenir un portrait global de l'effet de chaque variable explicative. L'analyse de plusieurs milliers d'enregistrements était déjà requise dans les études précédentes réalisées à partir des données IMIS, même si ces études ne portaient que sur un seul agent chimique à la fois (Lavoue et coll., 2013). L'analyse de 90% du contenu de la banque IMIS (c.-à-d. >500 000 enregistrements) représentait par conséquent un énorme défi de synthèse pour cette recherche doctorale.

L'approche de méta-analyse a donc été utilisée dans l'analyse principale du chapitre 3 et l'analyse secondaire du chapitre 4, ce qui a permis de dégager les tendances globales présentes dans la banque IMIS et de les présenter sous forme de paramètres de synthèse. Cette approche est donc particulièrement d'intérêt dans l'étude de sources de données multi-agents de taille élevée telles que les BDEP. De plus, les forest plots ont permis de positionner les résultats de chaque agent par rapport au résultat combiné et d'ainsi présenter visuellement l'ampleur de l'hétérogénéité des effets plutôt que de s'appuyer sur une simple statistique de test. Certains auteurs ont d'ailleurs discuté de la faible puissance du test d'hétérogénéité habituellement utilisé dans les méta-analyses (Hedges et Pigott, 2001; Borenstein et coll., 2010) et favorisent l'utilisation d'approches plus transparentes. Conceptuellement, les associations (ou leur absence) observées à travers une majorité d'agents chimiques suggèrent des tendances sous-jacentes présentes dans la source de données. En revanche, les associations variant d'un agent chimique à l'autre sont plus difficiles à interpréter et peuvent

résulter de pratiques spécifiques pour un agent en particulier ou être reliées à des facteurs non présents dans l'analyse.

L'utilisation de la méta-analyse requiert que les différentes unités statistiques (dans notre cas, les agents chimiques) de la procédure soient considérées comme des populations indépendantes. La situation extrême serait celle où l'ensemble des agents chimiques auraient été mesurés au même moment et au même endroit au cours des 67 000 visites enregistrées dans la banque IMIS. Une telle situation serait potentiellement associée à une forte corrélation entre les niveaux mesurés à travers les agents et pourrait fausser la méta-analyse. Cette corrélation ne serait néanmoins pas automatique car il faudrait également supposer que le processus de contamination et d'émission est le même pour tous les agents, ce qui est peu vraisemblable. Dans notre analyse, une médiane de 1 agent était mesurée lors d'une visite (fourchette entre 1 et 30). De plus, une analyse qualitative de la variable titre d'emploi (variable contenant des informations non standardisées) indique que lorsque plusieurs agents sont prélevés lors d'une visite, les échantillons sont habituellement recueillis dans des zones différentes. Certains agents sont parfois analysés sur un même milieu collecteur, mais plutôt pour des raisons de facilité analytique que pour refléter des situations d'exposition conjointes. On ne peut nier qu'une certaine corrélation puisse exister dans certains cas, par exemple entre des mesures de différents agents prises dans la même usine « sale », ou encore lorsque deux solvants fréquemment rencontrés dans les mêmes formulations sont mesurés durant la même visite. Les éléments précédents tendent cependant à démontrer que ce phénomène est minoritaire et que, de manière générale, les agents chimiques peuvent être considérés comme des populations indépendantes et différentes permettant l'utilisation de la méta-analyse dans notre étude.

Pour le manuscrit du chapitre 3, publié dans la revue *Annals of Occupational Hygiene*, les résultats d'analyse individuels pour chaque agent chimique ont été inclus sous forme de forest plots dans les annexes en ligne de la revue. Les images de tous les graphiques avaient préalablement été converties et regroupées dans un même fichier .html, ce qui facilite grandement la consultation des nombreux résultats directement sur le site web de la revue à partir d'un fureteur.

6.1.3 Prédiction du type d'exposition des résultats ND dans IMIS

Le troisième volet de la recherche doctorale, présentée au chapitre 5, visait à aborder un problème relié à la structure particulière de la banque IMIS : le statut d'une mesure codée ND est indiqué dans la même variable identifiant si l'échantillon correspond à une mesure courte durée ou VEMP-8h. Ceci complique leur interprétation puisque l'absence d'information sur la durée de la mesure et la valeur de censure empêche un utilisateur de traiter adéquatement un résultat ND. Les travaux récents sur IMIS (Hamm et Burstyn, 2011; Lavoue et coll., 2011; Lavoue et coll., 2013; Cowan et coll., 2015; Lee et coll., 2015) ont tous souligné les limites associées à l'absence de valeurs de censure pour les données ND. Cette absence d'information a forcé ces auteurs à les exclure de leurs analyses ou à transformer la variable réponse en simple dichotomie (inférieur/supérieur à un seuil, en général la VLE). Le fait de ne pas prendre en considération la durée de la mesure dans des modèles de prédiction pourrait fausser les estimés d'exposition réalisés à partir d'IMIS.

Les modèles CART ont été utilisés puisqu'ils permettent de modéliser des associations non-linéaires complexes entre les variables prédictives et sont facilement interprétables. Une approche transparente et reproductible a donc été proposée permettant 1) de développer ces

modèles CART pour chaque agent chimique en se basant sur les variables communes aux banques IMIS et CEHD, et 2) d'évaluer leur performance à prédire, parmi les résultats ND de la banque IMIS, lesquels correspondent à des mesures courte durée ou VEMP-8h.

Notre étude a permis de mettre en évidence que la proportion des mesures VEMP-8h était généralement plus faible (différence médiane de 18%) dans les données ND que dans les données détectées pour la partie commune aux deux banques. Ceci tend à démontrer que l'utilisation de méthodes d'attribution basées sur la proportion des catégories VEMP-8h/courte durée dans les données détectées (p. ex. attribution aléatoire, attribution de la catégorie la plus fréquente) serait inappropriée. Nos résultats ont d'ailleurs montré une meilleure performance des modèles CART comparée à ces deux méthodes d'attribution pour les agents chimiques les plus pertinents (c.-à-d. ceux ayant une proportion substantielle de mesures ND, courte durée et VEMP-8h – solvants et gaz). Quant aux agents ayant une forte majorité de mesures courte durée ou VEMP-8h parmi leurs mesures détectées, l'attribution de la catégorie la plus fréquente serait certainement appropriée. On note dans cette catégorie la plupart des métaux (mesures VEMP-8h) et des isocyanates (mesures courte durée).

Les résultats encourageants de ce volet de recherche ont permis de contribuer à l'avancement des connaissances sur la question méthodologique des données non détectées, traditionnellement rarement prise en compte. Les résultats de prédiction des modèles CART pourront être utilisés dans toute étude visant à produire des portraits d'exposition à partir des données IMIS et faciliteront l'attribution d'une valeur de censure aux données ND.

6.2 Limites de la recherche

Bien que les limites spécifiques à chaque manuscrit soient mentionnées dans les chapitres correspondant, des limites de type plus générales sont présentées dans cette section.

Premièrement, la portée de nos résultats est limitée par l'absence d'une source de données représentant un échantillon aléatoire et représentatif des travailleurs dans les milieux de travail aux États-Unis (c.-à-d. gold standard) sur la base de laquelle les données IMIS pouvaient être comparées. Bien que l'analyse des variables internes dans IMIS soit informative, cette approche ne permet pas d'établir directement si les données IMIS sont représentatives de la distribution réelle des niveaux expositions retrouvés dans la population générale. Plusieurs des variables incluses dans nos analyses étaient par ailleurs reliées à la stratégie de sélection des milieux de travail. Les prédictions effectuées à partir des données IMIS demeurent donc sujettes au jugement porté par l'utilisateur. Il est toutefois possible d'évaluer la sensibilité des résultats à ce type de jugement en testant différents scénarios de prédictions de niveaux d'exposition à l'aide de la modélisation statistique. Même si cette approche ne permet pas de rendre les données représentatives, elle permet de maitriser l'influence des différentes variables considérées sur les prédictions, pour rendre leurs valeurs plus proche de la population (p. ex. taille des entreprises) ou pour choisir des combinaisons jugées plus représentatives (p. ex. type d'inspection « planifiée »).

Deuxièmement, les évaluations quantitatives présentes dans IMIS résultent d'un calcul de concentration moyenne pondéré sur la base de résultats analytiques (p. ex. calcul d'une VEMP-8h à partir de plusieurs échantillons consécutifs de courte durée). Ces évaluations incluent donc une part de jugement professionnel. Bien que cette caractéristique puisse être considérée comme une valeur ajoutée aux données brutes (p. ex. connaissance du milieu évalué), les variations des pratiques d'estimation des concentrations moyennes pondérées

dans le temps et selon les régions compliquent leur interprétation au niveau populationnel. Les estimés d'exposition à partir des données IMIS pourraient donc être sujets aux différences de distribution de ces pratiques. À titre d'exemple, une différence de niveaux d'exposition observés entre deux régions pour un agent chimique pourrait être attribuable à des approches différentes de calcul d'une VEMP-8h entre ces régions. Même si les données brutes correspondant à l'évaluation pondérée étaient disponibles dans la banque CEHD, il n'a pas été possible d'établir avec certitude l'approche préconisée par l'inspecteur pour leur calcul en raison du manque d'information entourant ces données.

La troisième limite concerne l'utilisation de CEHD dans les analyses comparatives effectuées aux chapitres 4 et 5. Les associations et prédictions rapportées dans ces chapitres ont été estimées malgré les limites inhérentes à l'utilisation de la banque CEHD qui contient presque exclusivement les résultats d'analyse des inspections menées sous un régime OSHA fédéral. Considérant la proportion élevée de résultats d'exposition provenant d'inspections menées sous un régime OSHA d'État dans la banque IMIS (31%), la disponibilité de cette partie « étatique » des résultats d'analyse représente donc un élément clé dans l'évaluation de la portée des conclusions présentées aux chapitres 4 et 5.

Finalement, la quatrième limite concerne le manque d'information entourant les données d'exposition. Parmi les principales BDEP, IMIS est certainement considérée comme l'une des banques où la quantité d'information disponible est la plus faible. Le manque d'information sur le type d'emploi occupé par le travailleur et la nature des opérations effectuées limite particulièrement l'interprétation et la caractérisation des niveaux d'exposition. De plus, la forte majorité des prédicteurs considérés dans cette recherche doctorale étaient au niveau de l'inspection et avaient donc la même valeur pour l'ensemble

des mesures recueillies lors d'une même visite. La performance des modèles CART a certainement été entachée puisque ces modèles ne pouvaient qu'attribuer l'une des classes de type d'exposition (c.-à-d. VEMP-8h, courte durée) à tous les résultats ND d'une même inspection (les 7 prédicteurs dans cette étude étaient au niveau de l'inspection). Des recommandations sur les informations accompagnant les mesures d'expositions dans les banques de données ont été formulées par plusieurs groupes de recherche (Gomez, 1993; Harris, 1995; Lippmann, 1995; Rajan et coll., 1997; LaMontagne et coll., 2002) afin de faciliter leur utilisation future à des fins de recherche. Plusieurs de ces recommandations ont été appliquées dans le cadre de la réorganisation de la banque française COLCHIC (Vincent et Jeandel, 2001). Les auteurs ont souligné la nécessité de simplifier la saisie des informations et de permettre au personnel générant les données d'obtenir un retour sur l'information qu'ils ont enregistrée pour améliorer les taux d'enregistrement et minimiser les erreurs de saisie. COLCHIC inclut notamment l'information sur la stratégie de mesure utilisée, la nature de la tâche effectuée et la présence de moyens de contrôle de l'exposition.

6.3 Originalité de la recherche

Les travaux de cette thèse ont contribué au développement de connaissances autant sur la source de données d'exposition IMIS que sur le plan méthodologique. Pour la première fois, une BDEP a été analysée dans sa globalité, alors que 90% des données de la banque IMIS ont été caractérisées au moyen d'un même protocole d'analyse. Ceci représentait un important défi de synthèse requérant l'adaptation d'une approche de méta-analyse importée du domaine de l'épidémiologie. Au final, cette approche a permis d'obtenir un portrait général de l'effet des prédicteurs et de dégager les tendances globales présentes dans la banque IMIS. La première étude systématique sur le sous-enregistrement a également été réalisée dans le cadre

de cette thèse grâce à l'utilisation de la banque de données d'analyse CEHD. Une approche a également été proposée permettant de faciliter l'interprétation des résultats non détectés de la banque IMIS.

6.4 Perspectives et recommandations

Bien que ne permettant pas d'établir directement la représentativité de la banque IMIS par rapport à la distribution réelle des niveaux d'exposition retrouvés dans la population générale, il est souhaitable que les connaissances développées dans cette thèse soient considérées lorsque les données IMIS sont utilisées pour établir le portrait d'un agent chimique en particulier. Une attention particulière devrait être portée aux variables étant associées aux niveaux d'exposition à travers tous les agents puisqu'elles suggèrent des tendances sous-jacentes présentes dans IMIS. Ces variables devraient être incluses dans des outils prédictifs de l'exposition lors de l'estimation de l'exposition pour un agent en particulier. De plus, les résultats de prédiction des modèles CART pourraient être utilisés dans toute étude visant à produire des portraits d'exposition à partir des données IMIS puisqu'ils facilitent l'attribution d'une valeur de censure aux données ND. L'intégration de ces informations contribuerait à raffiner les estimations réalisées à partir des données IMIS et ainsi améliorer les pratiques nécessitant d'évaluer l'exposition. Celles-ci incluent notamment la surveillance de l'exposition pour identifier des secteurs où les niveaux d'exposition sont encore élevés, des projets d'évaluation d'impact de la mise en place de nouvelles valeurs limites, des outils de promotion de l'amélioration des pratiques de mesures et l'évaluation rétrospective de l'exposition dans des études épidémiologique.

Dans une perspective de soutien à l'utilisation appropriée des données IMIS, de nouvelles études devraient être réalisées pour explorer les différences entre la banque IMIS et d'autres sources de données accessibles au public afin d'évaluer sa représentativité par rapport à la population globale. De plus, les efforts de recherche visant à la création de titres d'emploi standardisés à partir des descriptions sommaires fournies dans IMIS devraient être poursuivis considérant l'importance du type d'emploi comme déterminant de l'exposition. Des recherches méthodologiques visant à améliorer l'interprétation des résultats ND dans IMIS sont également nécessaires. L'avancée récente de nouvelles techniques de modélisation, permettant d'estimer simultanément la prévalence de l'exposition et les niveaux moyens lorsque l'exposition est présente, représente une avenue prometteuse pour le traitement de cette question (Taylor et coll., 2001; Chu et Nie, 2005; Chu et coll., 2008). Il est également souhaitable que l'agence OSHA fournisse un processus de saisie de données plus standardisé pour les inspecteurs afin de s'assurer que l'ensemble de l'information sur les déterminants de l'exposition (p. ex. information sur le port d'équipement de protection individuelle, titre d'emploi, tâche effectuée) et sur la valeur de censure des résultats ND soit systématiquement enregistrée dans IMIS. Finalement, il est essentiel de rassembler les sources de données contenant les résultats analytiques de laboratoire effectués dans les laboratoires OSHA d'État afin de faciliter la compréhension des mécanismes d'enregistrements des résultats d'exposition dans IMIS et l'interprétation des résultats ND.

6.5 Conclusion générale

Cette thèse a permis de mettre en évidence plusieurs défis reliés à l'utilisation des données de contamination de l'air recueillies par l'agence fédérale américaine OSHA pour l'estimation des expositions professionnelles. Nous croyons que ces obstacles ne devraient pas dissuader

les chercheurs d'utiliser les données IMIS, d'autant plus que la plupart des sources d'informations sur l'exposition sont entachées de problèmes similaires. Le besoin croissant de sources de données représentatives sur l'exposition pour identifier les situations les plus à risque et ultimement prévenir le développement de maladies professionnelles incite à poursuivre l'évaluation du contenu de la banque IMIS.

Bibliographie

ACGIH. (2014) 2014 TLVs and BEIs. ACGIH. 978-1-607260-72-1

Bigert C, Gustavsson P, Straif K et coll. (2015) Lung cancer risk among cooks when accounting for tobacco smoking: a pooled analysis of case-control studies from Europe, Canada, New Zealand, and China. J Occup Environ Med; 57: 202-9.

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

Buckland ST, Burnham KP, Augustin NH. (1997) Model selection: an integral part of inference. Biometrics; 53: 603-18.

Burnham KP, Anderson DR. (2002) Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. New York, NY: Springer. ISBN-13: 978-1441929730.

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

Burstyn I, Teschke K. (1999) Studying the determinants of exposure: a review of methods. Am Ind Hyg Assoc J; 60: 57-72.

Chu H, Nie L. (2005) A note on comparing exposure data to a regulatory limit in the presence of unexposed and a limit of detection. Biom J; 47: 880-7.

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.

Cowan DM, Cheng TJ, Ground M et coll. (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.

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

Friesen MC, Coble JB, Lu W et coll. (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.

Friesen MC, Demers PA, Spinelli JJ et coll. (2006) From expert-based to quantitative retrospective exposure assessment at a Soderberg aluminum smelter. Ann Occup Hyg; 50: 359-70.

Friesen MC, Pronk A, Wheeler DC et coll. (2013) Comparison of algorithm-based estimates of occupational diesel exhaust exposure to those of multiple independent raters in a population-based case-control study. Ann Occup Hyg; 57: 470-81.

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 coll. (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. Appl Occup Environ Hyg; 8: 768-74.

Gomez 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.

Hall AH, Teschke K, Davies H et coll. (2002) Exposure levels and determinants of softwood dust exposures in BC lumber mills, 1981-1997. AIHA J (Fairfax, Va); 63: 709-14.

Hall AL, Peters CE, Demers PA et coll. (2014) Exposed! Or not? The diminishing record of workplace exposure in Canada. Can J Public Health; 105: e214-7.

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

Harrel FEJ. (2001) Regression modeling strategies—with applications to linear models, logistic regression, and survival analysis. New York, NY: Springer.

Harris R. (1995) Guideline for Collection of Industrial Hygiene Exposure Assessment Data for Epidemiologic Use. Applied Occupational and Environmental Hygiene; 10: 311-16.

Hastie T, Tibshirani R. (1990) Generalized additive models. New York, NY: Chapman & Hall/CRC.

Hedges LV, Pigott TD. (2001) The power of statistical tests in meta-analysis. Psychol Methods; 6: 203-17.

Henn SA, Sussell AL, Li J et coll. (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 coll. (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.

Hewett P, Ganser GH. (2007) A comparison of several methods for analyzing censored data. Ann Occup Hyg; 51: 611-32.

Hornung RW, Reed LD. (1990) Estimation of Average Concentration in the Presence of Nondetectable Values. Applied Occupational and Environmental Hygiene; 5: 46-51.

Hosmer D, Lemeshow S. (2000) Applied Logistic Regression, Second Edition. John Wiley & Sons, Inc.

INRS. (2015a) Base de données Fibrex. http://www.inrs.fr/publications/bdd/fibrex.html (Accessed 15 december 2014).

INRS. (2015b) SOLVEX. http://www.inrs.fr/inrs-pub/inrs01.nsf/IntranetObject-accesParReference/Rubrique9g/$FILE/visu.html (Accessed 15 december 2014).

Jones C, Weld L, Gray W et coll. (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).

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

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

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

Lavoué J. (2006) Évaluation de l'exposition professionnelle au formaldéhyde à partir de sources de données préexistantes. Montréal, Québec: Université de Montréal. 279.

Lavoue J, Beaudry C, Goyer N et coll. (2005) Investigation of determinants of past and current exposures to formaldehyde in the reconstituted wood panel industry in Quebec. Ann Occup Hyg; 49: 587-602.

Lavoue J, Droz PO. (2009) Multimodel inference and multimodel averaging in empirical modeling of occupational exposure levels. Ann Occup Hyg; 53: 173-80.

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.

Lavoué J, Gérin M, Bégin D et coll. (2012) Valorisation des données d'exposition professionnelle mesurées au Québec depuis 1980 par les équipes du Réseau public québécois en santé au travail - Étude préliminaire. Montréal, QC: IRSST.

Lavoue J, Gerin 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.

Lee D, Lavoue J, Spinelli J et coll. (2015) Statistical modeling of occupational exposure to polycyclic aromatic hydrocarbons using OSHA data. J Occup Environ Hyg; 1-14.

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 coll. (1998) Surveillance of respirable crystalline silica dust using OSHA compliance data (1979-1995). Am J Ind Med; 34: 547-58.

Lippmann M. (1995) Exposure Data Needs in Risk Assessment and Risk Management: Database Information Needs. Applied Occupational and Environmental Hygiene; 10: 244-50.

Lubin JH, Colt JS, Camann D et coll. (2004) Epidemiologic evaluation of measurement data in the presence of detection limits. Environ Health Perspect; 112: 1691-6.

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: 378-83.

Mater G, Paris C, Lavoue J. (2016) Descriptive analysis and comparison of two French occupational exposure databases: COLCHIC and SCOLA. Am J Ind Med.

McNally K, Warren N, Fransman W et coll. (2014) Advanced REACH Tool: a Bayesian model for occupational exposure assessment. Ann Occup Hyg; 58: 551-65.

Melville R, Lippmann M. (2001) Influence of data elements in OSHA air sampling database on occupational exposure levels. Appl Occup Environ Hyg; 16: 884–99.

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).

Nieuwenhuijsen MJ. (2003) Exposure Assessment in Occupational and Environmental Epidemiology. Oxford medical publications. 9780198528616.

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

Olsen E, Laursen B, Vinzents PS. (1991) Bias and random errors in historical data of exposure to organic solvents. Am Ind Hyg Assoc J; 52: 204-11.

Olsson AC, Gustavsson P, Kromhout H et coll. (2011) Exposure to diesel motor exhaust and lung cancer risk in a pooled analysis from case-control studies in Europe and Canada. Am J Respir Crit Care Med; 183: 941-8.

OSHA. (2014) SIC Manual - detailed information for a specified SIC, Division, or Major Group. https://www.osha.gov/oshstats/index.html (Accessed 2 November 2014).

OSHA. (2015) Chemical Exposure Health Data. https://www.osha.gov/opengov/healthsamples.html (Accessed 2 August 2015).

Peters S, Kromhout H, Olsson AC et coll. (2012a) Occupational exposure to organic dust increases lung cancer risk in the general population. Thorax; 67: 111-6.

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

Peters S, Vermeulen R, Portengen L et coll. (2011) Modelling of occupational respirable crystalline silica exposure for quantitative exposure assessment in community-based case-control studies. J Environ Monit; 13: 3262-8.

Raftery AE, Madigan D, Hoeting JA. (1997) Bayesian Model Averaging for Linear Regression Models. J Am Stat Assoc; 92: 179-91.

Rajan B, Alesbury R, Carton B et coll. (1997) European proposal for core information for the storage and exchange of workplace exposure measurements on chemical agents. Appl Occup Environ Hyg; 12: 31-9.

Sauve JF, Beaudry C, Begin D et coll. (2012) Statistical modeling of crystalline silica exposure by trade in the construction industry using a database compiled from the literature. J Environ Monit; 14: 2512-20.

Sauve JF, Beaudry C, Begin D et coll. (2013) Silica exposure during construction activities: statistical modeling of task-based measurements from the literature. Ann Occup Hyg; 57: 432-43.

Scarselli A, Binazzi A, Di Marzio D. (2011) Occupational exposure levels to benzene in Italy: findings from a national database. Int Arch Occup Environ Health; 84: 617-25.

Scarselli A, Binazzi A, Marzio DD et coll. (2012) Hexavalent chromium compounds in the workplace: assessing the extent and magnitude of occupational exposure in Italy. J Occup Environ Hyg; 9: 398-407.

Scarselli A, Di Marzio D, Marinaccio A et coll. (2013) Assessment of work-related exposure to polycyclic aromatic hydrocarbons in Italy. Am J Ind Med; 56: 897-906.

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.

Schinkel J, Ritchie P, Goede H et coll. (2013) The Advanced REACH Tool (ART): incorporation of an exposure measurement database. Ann Occup Hyg; 57: 717-27.

Stewart PA, Rice C. (1990) A source of exposure data for occupational epidemiology studies. Appl Occup Environ Hyg; 5: 359-63.

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.

Taylor DJ, Kupper LL, Rappaport SM et coll. (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 coll. (1999) Exposures to wood dust in U.S. industries and occupations, 1979 to 1997. Am J Ind Med; 35: 581-9.

Tielemans E, Schneider T, Goede H et coll. (2008) Conceptual model for assessment of inhalation exposure: defining modifying factors. Ann Occup Hyg; 52: 577-86.

Tielemans E, Warren N, Fransman W et coll. (2011) Advanced REACH Tool (ART): overview of version 1.0 and research needs. Ann Occup Hyg; 55: 949-56.

Tielemans E, Warren N, Schneider T et coll. (2007) Tools for regulatory assessment of occupational exposure: development and challenges. J Expo Sci Environ Epidemiol; 17 Suppl 1: S72-80.

US Department of Labor. (2014) OSHA Enforcement Data. http://ogesdw.dol.gov/views/data_summary.php (Accessed 15 october 2014).

Van Hulst A, Roy-Gagnon MH, Gauvin L et coll. (2015) Identifying risk profiles for childhood obesity using recursive partitioning based on individual, familial, and neighborhood environment factors. Int J Behav Nutr Phys Act; 12: 17.

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

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 P, Carton B, Fjeldstad P et coll. (1995) Comparison of Exposure Measurements Stored in European Databases on Occupational Air Pollutants and Definition of Core Information. Applied Occupational and Environmental Hygiene; 10: 351-54.

Wheeler DC, Archer KJ, Burstyn I et coll. (2014) Comparison of Ordinal and Nominal Classification Trees to Predict Ordinal Expert-Based Occupational Exposure Estimates in a Case-Control Study. Ann Occup Hyg.

Wheeler DC, Burstyn I, Vermeulen R et coll. (2013) Inside the black box: starting to uncover the underlying decision rules used in a one-by-one expert assessment of occupational exposure in case-control studies. Occup Environ Med; 70: 203-10.

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.

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

Zou GY, Donner A. (2013) Extension of the modified Poisson regression model to prospective studies with correlated binary data. Stat Methods Med Res; 22: 661-70.

ANNEXE 1- Creating standard company identifiers in OSHA administrative database

Creating standard company identifiers in OSHA administrative database

Introduction

The objective of the project was to match establishment names from a large administrative data set (4.2M records) to a smaller more specific data set (66K records), both these tables corresponding to subsets of the Integrated Management Information System (IMIS), a much larger database from the Occupational Safety and Health Administration (OSHA). The large administrative data set contains all citations issued to companies by inspectors of OSHA for non-compliance with standards, while the smaller one contains occupational exposure levels measured by inspectors during their visit to a company. Each record in the citation database corresponds to one visit, while each record in the exposure database corresponds to one measurement made during a visit. The main aim of the linkage was to be able to establish the citation profile of any company in particular, ultimately to study whether companies with high numbers of citations are associated to more elevated exposure levels. For the rest of the document we will refer to the large table as OSHA citation database, and call the smaller table OSHA exposure database.

The problem is that there is no standard unique company/establishment identifier in OSHA databases: company name, street address, ZIP code and state are provided in different fields in free text format. So the same company might be given slightly different names in different visits (e.g. ACME, inc., then ACME incorporated).

Based on the name, address, ZIP and state variables, we developed a technique that assigned a unique index to all the visits corresponding to the same establishment based on the

similarity of their names and location. This index can then be used as a standard identifier for establishment to link the citation and exposure databases.

Methods

Establishment vs. company: The unit of interest in our project is an establishment, which can be part of a larger company with several establishments in various locations. For this reason, to be regarded as corresponding to the same establishment, 2 records in the databases had to have the same ZIP and state. This means our algorithm will miss an establishment that moved over the years (if it changed ZIP code)

Link between the citation database and the exposure database: Company is not the only link between the citation and exposure datasets. The visit identifier is also present in both. This means for each unique visit in the exposure database, we could find data from the same visit in the citation database, and we knew this was the same company. Comparing company names in records linked through activity showed that names in the exposure database were truncated compared to the citation database. This might have happened during transfer into EXCEL of the exposure database at the time it was obtained through FOIA.

The creation of the unique company index was conducted as follows:

Each name in the exposure database was compared to other names in the same database with the same state and ZIP, and was assigned the same company index whenever they were within a 2-character difference. Because all names were compared to all others, it was possible that names with the same index might have a >2 character difference (A similar to B, and B similar to C implied A similar to C). In parallel, each name in the citation database (limited to names linked to the exposure dataset through their activity number) underwent the

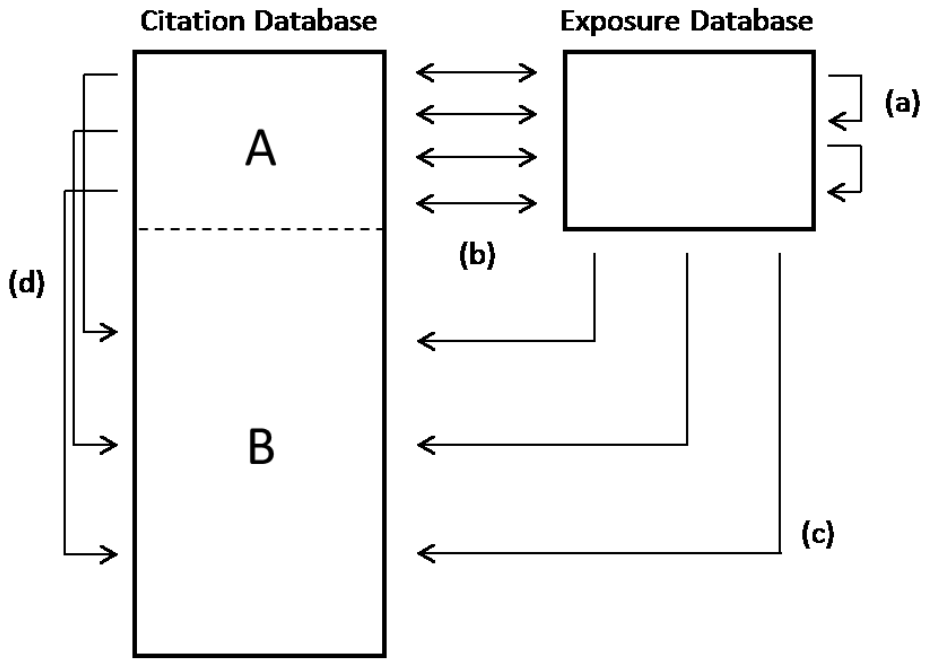
same procedure. Finally indices in the exposure and citation database were combined using the activity number link.

Before the search procedure was applied, we conducted preliminary cleaning of the names. First, we capitalized all the characters. Then, we removed a few common generic enterprise markers such as: INC., CORP., CO., MFG, etc. Finally, we also removed all non-alphanumerical characters such as: commas, periods, semi-columns, spaces, ampersands, brackets, etc.

When all the matching steps were completed, an index was assigned to all groups of matched and linked names.

An overview of the indexing procedure can be seen in figure 1.

Figure 1 : Summary overview of the indexing procedure between the citation and the exposure databases.



a) Names were matched within the exposure database (same ZIP and state)
b) Linkage between exposure and citation databases through activity number
c) Names were matched between the exposure database and part B of the citation database (same ZIP and state)
d) Names were matched between the part A and part B of the citation database (same ZIP and state)

Results

The exposure database contained 65 516 unique values of company names (combined with ZIP and State) without any treatment. Just applying the preliminary cleaning decreased this number to 64 800. After the full procedure was applied, 63 231 unique indices remained in the exposure database.

Table 1 below presents the distribution of the created indices according to the number of different company names associated with them in the exposure database, before and after the preliminary cleaning.

Table 1 : Distribution of company indices according to number of company names

Number of Unique Names in each Index (1)	1	2-3	4-10	> 10
Number of Indices - Original Names	43 933	17 110	2 181	7
Number of Indices - Clean names	48 262	14 019	948	2

(1) All the names used in this table are from the exposure database or they have been obtained from the citation database through a name from the exposure database.

Tables 2 and 3 describe the distribution of the company indices according to the maximal and minimal number of characters differences across all names associated with an index. The results are separated by dataset because of the issue of name truncation in the exposure database, which would create artificially large character differences. These results provide an idea of how 'far away' names could end up being associated by our procedure.

Table 2 : Distribution of company indices according to the maximal and minimal number of characters differences across all names associated with an index and specific to the exposure database.

Exposure Database				
Character Difference among the Names in each Index (1)	0	1-3	4-10	> 10
Number of Indices - Minimum Char. Diff.	62 741	303	147	10
Number of Indices – Maximum Char. Diff.	49 110	7 157	5 887	1038

(1) All the names used in this table are from the exposure database or they have been obtained from the citation database through a name from the exposure database. The values were calculated by using cleaned names (no punctuation).

Table 3 : Distribution of company indices according to the maximal and minimal number of characters differences across all names associated with an index and specific to the citation database.

Citation Database				
Character Difference among the Names in each Index (1)	0	1-3	4-10	> 10
Number of Indices - Minimum Char. Diff.	62 201	545	429	56
Number of Indices – Maximum Char. Diff.	61 129	877	994	231

(1) All the names used in this table are from the citation database. The values were calculated by using cleaned names (no punctuation).

In order to assess the quality of the grouping algorithms, we selected a random sample of indices for which all associated names were examined to decide whether the grouping was appropriate or not. This effort was stratified by database (exposure vs. citation database), and according to the maximum character difference (most likely mistakes would occur for large differences), with 100 indices per strata. The results of this validation are presented in table 4.

During manual investigation, if the grouping in an index was ambiguous, meaning that the names were not the same but could have possibly been referring to the same enterprise, the grouping of that index was flagged as "Undetermined".

Table 4: Distribution of a subset of company indices according to their grouping status.

	Exposure database		Citation database	
Maximum Character Difference among Names in each Index (1)	4 - 10	> 10	4 - 10	> 10
Successful Grouping	95	83	99	95
Failed Grouping	4	13	1	5
Undetermined Grouping Quality	1	4	0	0

(1) 400 random indices were selected: 100 from each "Maximum Char. Diff." subset based on tables 2 and 3. The "Undetermined Grouping Quality" row contains indices where its names were not obviously miss-grouped.

As we can see in table 4, the grouping is largely successful. It is unsurprising to see that the grouping quality is worse in indices carrying a large maximum character difference among their names. Nevertheless, the worse grouping percentage was at 83% in the exposure

database at > 10 character difference and at 95% for a 4 to 10 character difference. This could be extrapolated to indicate that, in the worst case, there would be a 0.76% error rate. The successful grouping rate would then be at 99.2%.

Two examples of erroneous groupings can be seen in table 5.

Table 5 : Example of erroneous grouping of similar names performed by matching algorithm. Characters in grey and spaces were removed before matching.

Enterprise Name	State	ZIP	Index
P C M.	MA	2780	363
CG MFG INC	MA	2780	363
B&J MANUFACTURING CORPORATION	MA	2780	363
C G MANUFACTURING, INC.	MA	2780	363
B&J MANUFACTURING CORPORATION	MA	2780	363
G G MFG INC	MA	2780	363
B&J MANUFACTURING CORPORATION	MA	2780	363
B&J MANUFACTURING CORPORATION	MA	2780	363
B&J MANUFACTURING CORPORATION	MA	2780	363
C G MANUFACTURING INC	MA	2780	363
C G MANUFACTURING INC	MA	2780	363

Enterprise Name	State	ZIP	Index
GAP MANUFACTURING, INC.	GA	30568	9545
GAP MANUFACTURING, INC.	GA	30568	9545
GMW INC.	GA	30568	9545
GAP MANUFACTURING, INC.	GA	30568	9545
GAP MANUFACTURING, INC.	GA	30568	9545

From the manual investigation, it became clear that a large number of miss-groupings occurred when the significant section of the name (disregarding generic industry markers: INC., CORP., etc.) was short. If the name only contained 2 or 3 letters, it was easily grouped with other short names. This is due to our 2 character difference threshold. In order to solve this issue, it would be possible to add extra steps to account for different name lengths. However, adding varying thresholds could also influence grouping of well grouped names.

This extra step could only improve the grouping of 0.76% of indices but could also disrupt others. We decided to accept this margin of error since the present result is largely satisfying.