A Feasibility Study for Developing an Occupational Exposure-Control Intelligence System in Great Britain (using Respirable Crystalline Silica as the Working Example)

Final Report

Thomas Ashton Institute 17 December 2021 @ashton.manchester.ac.uk

Authors

Ioannis Basinas[1] PhD, Julia Rozanova[2] MSc, Andre Freitas[2] PhD,

Martie van Tongeren[1] PhD & Damien McElvenny[1] PhD

[1]Thomas Ashton Institute for Risk and Regulatory Research & Centre for Occupational and Environmental Health, Epidemiology and Public Health Group, Division of Population Health, Health Services Research and Primary Care; School of Health Sciences, Faculty of Biology, Medicine and Health

[2]Department of Computer Science, School of Engineering, Faculty of Science and Engineering

University of Manchester

Funded by the Health and Safety Executive.

Contents

Key Messages

Occupational respiratory diseases are the subject of one of the Health and Safety Executive's (HSE) Health Priority Plans. There are an estimated 12,000 deaths from these diseases each year. It includes a wide range of conditions, some of which develop a short time after exposure (e.g. asthma, legionella infections) and others many years later (e.g. pneumoconiosis, lung cancer). Estimated trends in exposures and exposure controls are leading indicators of what the future burden of work-related lung diseases might be.

For many years, the HSE in Great Britain has maintained a National Exposure Database (NEDB). The database contains results from workplace exposure measurements on hazardous substances collected by HSE and by industry. The volume and variety of exposure data collected was much higher in the 1980s and 1990s than in the last 2 decades. The original aims of NEDB included using the data to inform policy-making and standard-setting bodies about workplace exposure levels in Britain. However, the data have fallen short of fully meeting these aims in recent years. This report examines the feasibility of establishing an occupational exposure-control intelligence system (OccECIS) which will bring together existing data on workplace exposure and control measures, to provide on-going data analysis and reporting on leading indicators related to agents that have the potential to cause occupational respiratory diseases. The report used respirable crystalline silica (RCS) as a working example.

The main findings were:

- The review of data sources suggested that good employment data regarding industries and occupations are available from GB national data sources (e.g. census). Established indicators for the prevalence of exposure among certain occupations and/or industries are also available for some respiratory health related agents e.g. in the form of job exposure matrices.

- Similarly, data on exposure levels to respiratory agents, such as respirable crystalline silica exists and are being collected by various stakeholders (industry, consultants, researchers, HSE).

  The data may not be readily available, so negotiation with and/or incentivisation of the data holders will be required.

- Comprehensive and representative data on the use of control measures and their effectiveness are not currently available; specific surveys will be required to collect such data for inclusion in any future system.

- The review of exposure-control intelligence systems demonstrated that a system similar in scope to OccECIS does not currently exist elsewhere.

- A conceptual framework for OccECIS has been developed.

- A series of theoretical questions and a gap analysis were used to inform technical requirements for an OccECIS.

- Feasible technical solutions have been proposed for data extraction, data storage and intelligent modelling (including addressing issues of bias, uncertainty and data quality) of occupational exposure and control data.

- An assessment of HSE's Exposure Control Indicator (ECI) data is required in order to determine to what extent further data collection on exposure control information will be required.

- It will not be possible to establish a fully automated OccECIS, because expertise in occupational exposure assessment and mathematical/statistical modelling will be required to accommodate data gaps and ensure data and analyses are appropriately interpreted.

Executive Summary

The overall aim of this project was to assess the feasibility of developing an occupational exposure control intelligence system (OccECIS) that could be used to provide data on leading indicators (for occupational lung diseases). The HSE and its stakeholders could use this system to prioritise hazards and sectors and occupations of concern, in order to inform future interventions to reduce exposures and to monitor the effectiveness of such interventions over time and so limiting the future burden of occupational lung diseases. Data for respirable crystalline silica (RCS) was used as the working example (as this is a recognised priority for HSE). The feasibility questions with a summary of their answers are set out below.

To describe the available data on agents that cause work-related respiratory diseases

We have identified a number of data sources. It is acknowledged that a large amount of data are available, although much will be of a historical nature going back 30 to 40 years and important data gaps exist for current exposure levels. We would recommend that data going back no more than about 20-25 years should be included in OccECIS.

To determine whether the required data sources are available

For RCS and the other agents that cause work-related respiratory diseases, we believe it is feasible to establish lists of reliable sources of exposure data. However, these data are likely to be unevenly spread across exposure scenarios, and given resource requirements, some prioritisation for filling the data gaps will be required. Any data collection should ensure that data are of sufficient quantity and quality and can be organised into a format compatible with that chosen for OccECIS. Some data may not be publicly available or may require funding to access. Robust procedures will need to be developed to combine individual with aggregated exposure data. For data on risk management measures (RMM), including both prevalence and impact on exposure levels at a population level, there is limited, information available. Support of an OccECIS with new exposure data collection initiatives will be an important component.

To describe data gaps on occupational exposure to substances, in terms of prevalence and intensity

Both macro (e.g. substance, industry) and micro-scale (e.g. individual occupations/ processes, periods of coverage) analyses of the data gaps are required. We have (partly) carried out such an analysis for silica, but it should be undertaken for other

priority agents. Criteria will need to be developed to identify the priority data gaps that need to be addressed for the system to fulfil its purpose. This will involve numbers estimated to be exposed, level of exposure, and feasibility of implementing risk management measures.

To determine the available intelligence on what risk management methods are in place in different sectors or occupations to control or reduce exposure levels

Our gap analyses suggest that the available data related to the prevalence of specific RMMs in British or any other country's workplaces are rather limited. This is in contrast with the efficiency of different exposure control measures where several dedicated systems and databases are available, ranging from generic ones to systems specific to certain exposures and/or industries. Targeted data collection exercises will therefore be required to cover this lack of intelligence regarding the presence and prevalence of RMMs. Any future efforts to collect these data should include elements that will allow the periodic update of the information held by the system in its database so as to ensure that the evaluation of potential intervention efforts are properly supported.

To determine how different types of data can be captured most efficiently and integrated into the database

Clearly, some data capture methods are more efficient than others are. We would propose that systems are developed for capturing structured and unstructured data that exists either in aggregated or summary form or as individual exposure measurement. It is important that contextual information is captured also. We believe that occupational hygiene and statistical expertise are required to appropriately capture issues of data quality, bias and uncertainty. We recommend negotiating access to structured database application programme interfaces (APIs) wherever possible, investing in the development of specialized information extraction algorithms where PDF-style data sources are sufficiently regular in structure. We emphasise the strong need for a data curator role that can maintain the automated data flow and recognize when changes need to be made, or where manual extraction/input is the only option for highly unstructured new inputs. This includes the identification of exposed groups (i.e. occupation, industries), tasks and processes as well as the estimation and quality control of the estimates for the proportions of exposed workers whenever required.

To determine how to make the system dynamic and easily updatable

There needs to be defined minimum data standards for which any of the OccECIS proposed standard outputs can be updated using appropriate mathematical and statistical models that should be developed. However, for most situations or scenarios, it is likely that appropriate judgement will need to be made about any new data added to the system (quality, bias, uncertainty) and how it is related to existing data in the system. In some circumstances where the data are more descriptive than numerical, then qualitative approaches may need to be employed. Expert hygiene and mathematical or statistical judgment may be required. The data curator role would also interact with domain experts on suggested updates to the system and integration of new data sources. For most cases, however, data input should be encouraged through a structured front-end service (in collaboration with domain expert assessors) supported by strong incentive structures and good relations with industry stakeholders.

To define how exposure-control data can be analysed and exposure prevalence and intensity (high/medium/low) be determined by sector/industry, occupation, age and time period

Exposure intensity will be categorised in general terms. For substances for which workplace exposure limits are available defined cut-off levels for high, medium and low exposures can be established. Alternative approaches are also available, and the final approach chosen may need to be tailored to the specific substances or exposure circumstances. Some workers may experience exposures that are highly variable between and within working days and monitoring may be performed under worst-case scenario approaches, which may lead to results that are not representative of daily working exposures. To reduce this potential the occupational groups of interest could be assigned an exposure level following an analysis of the exposure distribution within the measurements available. The level can then be assigned on the basis of a defined proportion of the available measurements that exceed the chosen cut-off limit. Alternatively, information related to the between and within worker variation could be integrated to weight the mean estimates and properly assign the group to an exposure category. Such an approach of course will need to further account for the absence of measurement data and/or the presence of non-GB data on any sample and expert opinion could be utilised in these cases. In the absence of workplace exposure limits (WELs) an appropriate percentile of the exposure distribution could be utilised.

To make recommendations on how data quality should be assessed

We propose that standardised techniques be developed and used for assessing data but also output quality. In principle, for measurement data, quality can be evaluated on the basis of standard approaches accounting for limitations in the methods used to collect the data, the sampling strategy applied, and the completeness in terms of contextual information reported. Information availability is inherently associated with the study aim and OccECIS is a system that will serve multiple aims. It is thereby essential that data are not excluded a priori on the basis of missing information. Instead, a core set of key contextual information that need be available to define a minimum level of quality for data to remain in the database will need be established -

this is essential in defining some of the future uses of the system. Implemented sampling and analytical methodologies will need be considered in terms of their adequacy alongside the representativeness of the results in terms of the variability present in the workplace. For those data included in the system, quality criteria will also need to be developed to enable appropriate interpretations to be associated with any analyses derived by the system.

To propose a methodology for how will trends over time might be assessed

Here we need distinguish between two types of trends - exposure prevalence and exposure intensity. For the first of these, trends in number of workers in an occupational or industry can be assessed using data from official statistics. These included data from the Census that takes place every 10 years, in combination with the Labour Force Survey or other relevant national surveys. For the second, for substances for which sufficient measurement data are available, classical approaches of modelling trends could be employed. This involves the application of linear mixed effect regression analysis, and/or of general additive modelling, to examine patterns of change in exposure levels across years or well-defined time periods. For substances for which insufficient measurement data are available, historical trends in exposure could be determined using expert statistical modelling approaches (expert-crafted Bayesian reasoning approaches) or by read across approaches. Further work may need to be carried out to assess if these trends have persisted.

To determine plausible uncertainty ranges for exposure-control prevalence and exposure-control trend estimates

Although some data in relation to the efficiency and surrounding uncertainties of different control measures is available in databases such as COMED, there is little information about their actual prevalence and working status or efficiency. Given the likely lack of data on RMMs, this question is probably the one that is hardest to respond to and for which feasibility is likely to be at its most problematic. Approaches may need to be developed on the use of RMMs and their general effectiveness. This could be in the form of worker surveys, through company surveys or through expert elicitation.

Conclusions

We believe that it is feasible to develop a system that can be used to monitor the effectiveness of any policy intervention. The quality of input data will largely determine the quality of the outputs of the system. In addition, i is important that such a system is not just a repository for historical data, but that it is continuously updated with new data collected by various stakeholders. Overall, we believe that it is feasible and important to develop an occupational exposure-control intelligence system in order: i) to derive estimates of exposure intensity and prevalence, as well as data on the use of control measures, which can subsequently be used so by HSE and other stakeholders; ii) to identify and prioritise hazards and sectors and occupations of concern; and iii) inform interventions to reducing the future burden of work-related respiratory diseases. The biggest barrier to establishing the system will be the incentivisation of those holding relevant data to contribute to this proposed national resource. A data curator role will need to be established to oversee the process of adding data to the system, including the quality control. We believe that developing a prototype system with RCS as the working example is the most appropriate initial task.

1. Introduction

Occupational respiratory diseases are the subject of one of the Health and Safety Executive's (HSE) Health Priority Plans. There are an estimated 12,000 deaths from these diseases each year. It includes a wide range of conditions, some of which develop a short time after exposure (e.g. asthma, legionella infections) and others many years later (e.g. pneumoconiosis, lung cancer). Estimated trends in exposures and exposure controls are leading indicators of what the future burden of work-related lung diseases might be.

For many years, the Health and Safety Executive (HSE) has maintained a National Exposure Database (NEDB). The database contains workplace exposure measurements collected by HSE and by industry. The volume and variety of exposure data collected was much higher in the 1980s and 1990s than in the last 2 decades (1). The original aims of NEDB include using the data to inform policy-making and standard-setting bodies about workplace exposure levels. However, the data have fallen short of fully meeting these aims in recent years. With that in mind, HSE commissioned the University of Manchester, via the Thomas Ashton Institute, to assess the feasibility of establishing a comprehensive occupational exposure control intelligence system (OccECIS) that can be used to record and monitor occupational exposure levels and effectiveness of controls in Great Britain.

The overall aim of this project was to assess the feasibility of developing an occupational exposure-control intelligence system (OccECIS), using respirable crystalline silica (RCS) as a worked example. The idea was that OccECIS would enable the HSE and its stakeholders to prioritise and implement intervention activities to control exposures towards agents that cause respiratory diseases in workplaces in Great Britain and to monitor their progress. If feasible, this could be extended to other health outcomes. The ultimate aim of OccECIS would be to provide a comprehensive assessment of occupational exposure levels (for both intensity and prevalence) for respiratory hazards and their changes over time as well as the use of risk management measures to control these exposures. These data could subsequently be used to predict in the impact of these exposure on the respiratory health of GB workforce and to prioritise any interventions.

For the OccECIS to be able to track exposures over time in scenarios and to evaluate the impact of both policy and practical interventions, it is important that OccECIS has the ability to be easily updated as a result of the inclusion of additional data.

Ultimately, the success of the OccECIS will be judged on its ability to produce leading indicators for the monitoring and evaluation of the potential impact of HSE's intervention activities on the prevention of occupational lung diseases.

The **primary** research objective addressed by this report was to answer the question:

- Is it feasible to develop a methodology for a British OccECIS that will provide intelligence on workplace exposures?

This is so that HSE and its stakeholders might in future develop leading indicators to:

- identify and prioritise hazards and relevant sectors and occupations of concern;
- establish appropriate interventions to reduce these exposures; and
- monitor the effectiveness of these interventions over time

The **secondary** objectives were:

- To describe and evaluate the available data on agents that cause work-related respiratory diseases.
- To describe data gaps for occupational exposure to substances, in terms of prevalence and intensity.
- To determine the available intelligence on what risk management methods are in place in different sectors or occupations to control or reduce exposure levels.
- To determine how different types of data can be captured most efficiently and integrated into the database.
- To determine how to make the system dynamic and easily updatable.
- To define how exposure-control data can be analysed and be classified into (high/medium/low) levels of exposure intensity and how exposure prevalence may be determined by sector/industry, occupation, age and time period.
- To make recommendations on how data quality should be assessed and the levels of exposure by high/medium/low should be assessed.
- To propose a methodology for how will trends over time might be assessed; and
- To determine plausible uncertainty ranges for exposure-control prevalence and exposure-control trend estimates.

This report describes how we have addressed these questions and includes a section on the conceptual specifications and the technical specifications. We have used the example of exposure to respirable crystalline silica to determine the availability of and access to relevant data, and to demonstrate what the outputs from the system may look like. Finally, we provide a summary of our findings in relation to the research questions and the overall feasibility of establishing an OccECIS and provide conclusions and recommendations.

2. Conceptual Specification

2.1. Methodology for Developing a Conceptual System Design

To develop the methodology for establishing the system, we first performed a scoping exercise to map previously established or ongoing occupational exposure systems and to map the availability of relevant exposure databases. To do this, we have built on previous work, part of an earlier study examining the feasibility of developing a surveillance system for monitoring trends in exposure to dangerous substances in EU workplaces performed as part of the EU-OSHA's Healthy Workers campaign (2). In addition, our information sources were further expanded through the links developed by the establishment of an international Expert Advisory Committee (EAC) (Table 1).

Table 1. Composition of the Expert Advisory Committee

Expert name	**Affiliation**
Tapani Tuomi	Finnish Institute of Occupational Health, Finland
Kelvin Williams	Kelvin Williams Ltd - Occupational Hygiene Consultancy, UK
Kevin Bampton	British Occupational Hygiene Society (BOHS), UK
Cheryl Peters	University of Calgary, Canada
Susan Peters	Utrecht University, Netherlands
Lin Fritsch	Curtin University, Australia
Amanda Eng	Massey University, New Zealand
John Cherrie	Institute of Occupational Medicine, UK
Lothar Lieck	EU-OSHA, Spain

Coincidentally, one member of the Advisory Committee (Dr Amanda Eng. Massey University, New Zealand) had been working on a similar project aiming to develop a national worker exposure database for New Zealand. She kindly shared her progress to date that covered intelligence related to the existence, content and availability of exposure data in national and international databases. These resources were further enriched through input from the other members in the EAC and by integrating additional information provided by the HSE project management team overseeing the project.

For demographic and employment data we carried out generic searches on the catalogues of the Office for National Statistics and its Data Service to identify the most relevant and detailed data in terms of resolution in time at a British and National level.

In addition, we identified databases and resources that were most relevant to occupational exposure to respirable crystalline silica (RSC).

The identified resources (i.e. databases and intelligence systems) were analysed in relation to their basic characteristics and relevance to OccECIS and together with input from the EAC were used to establish the basic conceptual framework upon which the system development should be based upon. This conceptual framework was then translated to a set of theoretical questions to be answered by the system in its outputs. These theoretical questions formed the basis for separate stakeholder and gap analysis upon the outputs of which the technical solutions for the system were based.

An overview of the methodology we applied to establish the general design of the OccECIS system is provided in Figure 1. Further details for each of the different steps described above and in the figure is provided in the sections that follow.

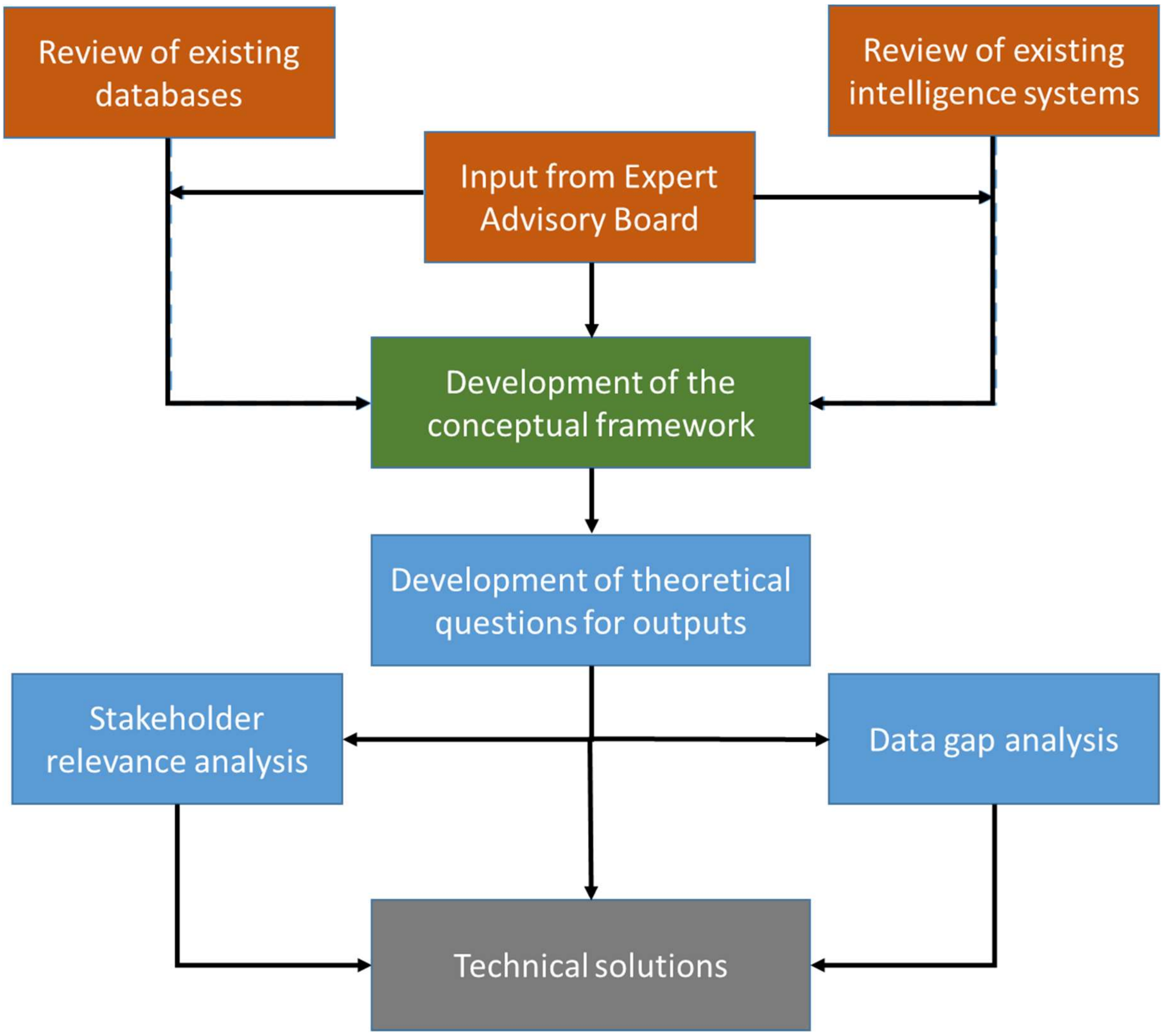
Review of existing databases
Review of existing intelligence systems
Input from Expert Advisory Board
Development of the conceptual framework
Development of theoretical questions for outputs
Stakeholder relevance analysis
Data gap analysis
Technical solutions

Figure 1. Methodology for identifying technical solution for OccECIS

2.2. Review of Existing Occupational Exposure-Control Intelligence Systems

As noted above, existing exposure-control intelligence systems were identified and reviewed through a hybrid approach combining personal knowledge with literature searches. This involved a scoping exercise that mapped existing systems based on a list of data sources already available (2) supplemented by input from the EAC and shared intelligence with colleagues at the Massey University, New Zealand.

These resulted in the development of an inventory of relevant occupational exposure intelligence systems. The inventory covered basic characteristics of the system in question such as country of relevance (if applicable), the type of information provided, the availability format, the name of the holder/management institution, and links to relevant references and information material. An overview of the main systems is provided in Table 2 with more details provided in Annex 1

Table 2. List of existing occupational exposure intelligence systems relevant to OccECIS

System name	Country	Coverage			Substances	Reference
		Intensity	Prevalence	RMMs*		
CARcinogen Exposure (CAREX) EU	EU	Yes	Yes	No	Carcinogens	(3)
CARcinogen Exposure (CAREX) CANADA	Canada	Yes	Yes	No	Carcinogens	(4)
Italian register of occupational exposures to carcinogen agents (SIREP)	Italy	Yes	Yes	No	Carcinogens	(5)
Exposure control efficacy library (ECEL)	NR	No	No	Yes	Any	(6)
CPWR's Exposure Control Database	NR	No	No	Yes	Silica, welding fumes, noise, lead	(7)
Silica Control Tool^{TM}	NR	No	No	Yes	Silica	(8)
Control Measures Efficacy Database (COMED)	NR	No	No	Yes	Any	(9)

*RMMs= Risk Management Measures (also known as Exposure Controls); NR = Not relevant

Several efforts to develop intelligence systems related to occupational exposures have been made in the past, mostly on an ad hoc basis towards fulfilling specific needs related to research around cancer. CAREX EU was initiated with the aim of estimating the numbers of workers exposed to (suspected) human carcinogens in the member states of the European Union (EU) (3). The effort was coordinated by the Finish Institute of Occupational Health (FIOH) and led to the establishment of an offline MS Access database that covered the period between 1990 and 1997. It included brief information on the substances and typical exposure conditions and provided cross industry and occupation specific exposure intensity and prevalence estimates. The CAREX EU system provided input for probabilistic investigations of the socioeconomic, health, and environmental impact associated with a range of policy options for amendments to Directive 2004/37/EC (Carcinogens or Mutagens at work) performed as part of the so-called “SHEcan” study (http://www.occupationalcancer.eu/) (10). Although comprehensive for its time, CAREX was based on a limited number of now “outdated” measurement data and there are currently no plans for maintenance and continuity of the system. Despite these limitations, CAREX EU pioneered the approach for future more systematic efforts by other countries and stakeholders to establish useful exposure systems.

One such effort and probably the most comprehensive to date, among the systems outlined in Table 2 is CAREX Canada (https://www.carexcanada.ca/). CAREX Canada aims to support Canadian legislative and public health organizations and the industry in prioritizing exposures to carcinogens in the workplace, and in developing targeted policies and programs to reduce exposures in the workplace. It covers 80 established or probable carcinogens in a relatively simple web interface that provides information on four different domains: a) substance characteristics, b) environmental exposure characteristics, c) occupational exposure characteristics and d) additional resources, which among others provide links to reading material containing information related to exposure reduction approaches. Data are sourced from national databases (e.g. Canadian Workplace Exposure Database, the Statistics Canada 2016 Census of Population), the literature and from the opinion of experts. Most of the system maintenance and data entry is done manually, and therefore is labour-intensive to maintain. Quality control, data review and extraction is also performed manually with several occupational hygienist and exposure scientists being involved in the process, which includes identifying and assigning levels of exposure to groups of workers. Given its importance and relevance to OccECIS, we have carried out a more in-depth analysis of this system. A brief summary of the results of this is shown in Table 3.

Table 3. Basic characteristics of CAREX Canada

Parameter	Output type	Output details	Input source	Method
Substance characteristics	Narrative	General info on substance, use, regulation (OELs), production and links to sources	- Literature, expert knowledge, own system results	Manual
Worker characteristics	Statistical	Exposure prevalence (no. and % of workers exposed) per occupation, industry, province and exposure level	- Census data for total number of workers in different strata - Literature reviews and expert opinion for identifying exposed groups	Manual entry and update, interfaced tool for detailed estimates
Exposure characteristics	Statistical	Average level per occupation and industry. Exposure category (low, medium, high)	- The Canadian Workplace Exposure Database (CWED) - Literature	Manual entry and update, interfaced tool for detailed estimates
Risk and case estimates	Narrative	NA	NA	NA
RMM efficiency/ interventions	Narrative	Links to external information sources	Guidelines, literature	Manual

NA= Not Available

Besides CAREX, other important systems identified include the Exposure Control Efficacy Library (ECEL) (6) and the Control Measures Efficacy Database (COMED) (9). Both systems aim to provide evidence on the efficacy of interventions to control inhalation exposure in workplaces. ECEL was developed in the late 2000's and was based on a comprehensive literature review on the effectiveness of risk management measures (RMMs). Initially, the system covered studies published between January 2000 and December 2007 and was available as an offline MS Access database. ECEL contains data on efficacy and related contextual data (e.g. industry, task, agent, exposure form, route, study type, location etc.) for six groups of risk management measures (i.e. enclosure, local exhaust ventilation, specialized ventilation, general ventilation, suppression techniques and separation of the worker) based on 90 peer reviewed publications (6). Further work on the system was performed in the late 2010’s. Besides an update of the underlying literature review with more recent data, a web-based interface was developed which included a semi-structured data-reporting format. The interface follows a semi-structured design with a rather simple web-interface where selection of the relevant entries is allowed on the basis of the included data fields. The database and system appear to accommodate only a manual updating procedure. ECEL is maintained and hosted by the Netherlands Organisation for Applied Scientific Research (TNO) (https://diamonds.tno.nl/ecel/risk-managements).

COMED is a more recent comparable initiative headed and hosted by the ITEM Fraunhofer Institute (https://comed.item.fraunhofer.de/) and supported financially and with expertise from the British Occupational Hygiene Society (BOHS) and was highlighted as an important development by the EAC. This system is at present in a development/optimization phase and thereby not yet fully available to the public. To gain better understanding of the system we contacted the developers in order to request further details as well as temporary access to the system. Overall, COMED contains a mixture of literature data combined with new measurement data provided by stakeholders (mainly occupational hygienists). A data entry platform has been developed for self-entry with the database adhering to a structured design where measurement data (with and without RMM and efficiency levels) can be entered and summarised alongside relevant contextual information (i.e. substance, process, sector of use, exposure duration, study design etc.) Once entered data are reviewed by experienced occupational hygienists and assigned a quality level on the basis of completion of the information required for the specific scenario. Queries through the interface allow visualisation of the data under specific scenarios (process, sub-process, RMM, data quality). Presently the system contains approximately 200 different scenarios mainly associated with welding. Developers are in the process of establishing collaborations and sourcing the required funding to further enrich their database.

Besides the above, semi-quantitative and quantitative general population Job Exposure Matrices (JEM) containing information on the prevalence and intensity of exposure to certain carcinogens across occupations and time periods comprise the simplest forms of such intelligence systems. Typically, the development of such JEMs is focused on a certain population or study and is based on a combination of expert evaluations and evidence from measurements. Good examples of such approaches include the Finnish job-exposure matrix (FINJEM)(11), the Nordic Job Exposure Matrix (NOCCA-JEM)(12), and/or the INTEROCC-JEM (13), An inventory of JEMs and other tools for exposure assessment is available through the EU-funded project OMEGA-NET (https://omeganetcohorts.eu/).

2.3. Review of Relevant Existing Data sources

Supported by the EAC we identified and then and reviewed relevant existing data sources. The process resulted in the identification of almost 60 exposure databases, and of several workforce and employment databases. A complete list of data sources is provided in Annex 1. Several of the identified exposure databases contained British or UK data – e.g. the National Exposure DataBase (NEDB) (14), SYNERGY (15), Combined DUST (CODUST) (16, 17), and WOOD Exposure (WOODEX) (18), although exposure data in these resources overlap as these were all to a large extent based on data from the NEDB.

Concerning workforce demographic data, several relevant databases are available with the most important being the UK census, which forms a near complete picture of the characteristics of the British population at a specific point of time (latest currently available census data are from 2011, although data from the 2021 census data will be available next year). Other surveys such as the Annual Population Survey (APS), which is a continuous household survey, provide estimates on social and labour market characteristics of the British population. These surveys provide very good resolution for the distributions of employed persons within industries and occupations. Demographics and area distribution statistics are also available with coverage in terms of time periods going back over 30 years at least.

Unfortunately, accessing the occupational exposure data is far from straightforward. Requirements for accessibility to the identified sources varies considerably depending on the reason the data have been collected and related underlying license as well as the policies of the holding institution. For example, most of the workforce demographic databases hold publically available data that can be accessed in a straightforward manner. For example, the APS is used to provide annual estimates of the total number of workers working across certain industries and jobs. Unfortunately, the projections for occupation and industry are not available at the desired resolution. Consequently, either accessibility needs be established through the secure servers of the Office for National Statistics (ONS) and the UK Data Service, which follows certain requirements[1] to calculate these projections, or a fee needs to be paid directly to the service to extract and process the required data itself[2]. The fee will depend on the resources required by the service to produce the requested dataset[3] given the dynamic nature of OccECIS that will need in this case to be provided on an annual basis. On the other hand, substance attribute databases

[1] https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/approvedresearcherscheme

[2] https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/makingarequest

[3] https://www.ons.gov.uk/aboutus/whatwedo/statistics/publicationscheme/chargingrates

are all “open access” available and thereby no delays or fees are expected to be paid for obtaining the underlying required data for OccECIS.

Accessibility to most of the exposure databases will be possible under a formal data sharing process as demonstrated in several earlier data pooling efforts (15-17, 19). Most of these databases are not held in the public domain and, generally, contain data that are owned and licensed for use by the holders, in which case institutional agreements for obtaining access to the data will be required. Previous experiences suggest that logistics, resources (working time and costs) and requirements for setting formal arrangements (e.g. data transfer and sharing agreements, protocols) combine to form significant obstacles in the data sharing process. Chargeable costs for gaining access may also be apply but at present this has not been explored in detail as contact with the data holders was not established as part of the present feasibility study.

Accessibility issues may apply, though likely to a lesser extent, also to previously established JEMs and intelligence systems. Public availability of such data sources also varies though access to most JEMs are unlikely to involve material costs. If this is not the case, then cost would be expected to be one-off and not prohibitive. For intelligence systems, most of the meta-data are usually accessible online (e.g. ECEL, COMED, CAREX Canada). Underlying data will naturally require permission for access and use from the original data holders. It is important to note that some of the most valuable ones, such as COMED, will require periodic updates to keep up-to-date and support the dynamic nature if attached to OccECIS. Thereby maintenance support from time to time may be required depending on the nature of relationship to be established.

2.4. Conceptual Framework for OccECIS Development

The collected information in relation to available resources were analysed alongside the input received by the EAC to establish the basic conceptual framework for the development of the system. A schematic summary of the established framework is provided in Figure 2. The system comprises of four domains surrounding the topics of: a) substance characteristics, b) demographics/worker characteristics, c) exposure intensity and d) risk management measures. Data relevant for each of these domains are contained in individual tables, which are subsequently linked to establish the system’s (OccECIS) database. The system’s database can subsequently be utilised to provide, for each relevant substance, the levels and prevalence of exposure, including estimates for different time periods, industries, jobs and tasks/processes. It will also describe the exposure conditions and control measures present. Brief description of potential sources that could be used as data inputs within the process as well as summaries of the potential system outputs are also provided. Naturally, the availability and relevance of this information will be different between substances with for example CAS numbers being irrelevant for process-generated substances.

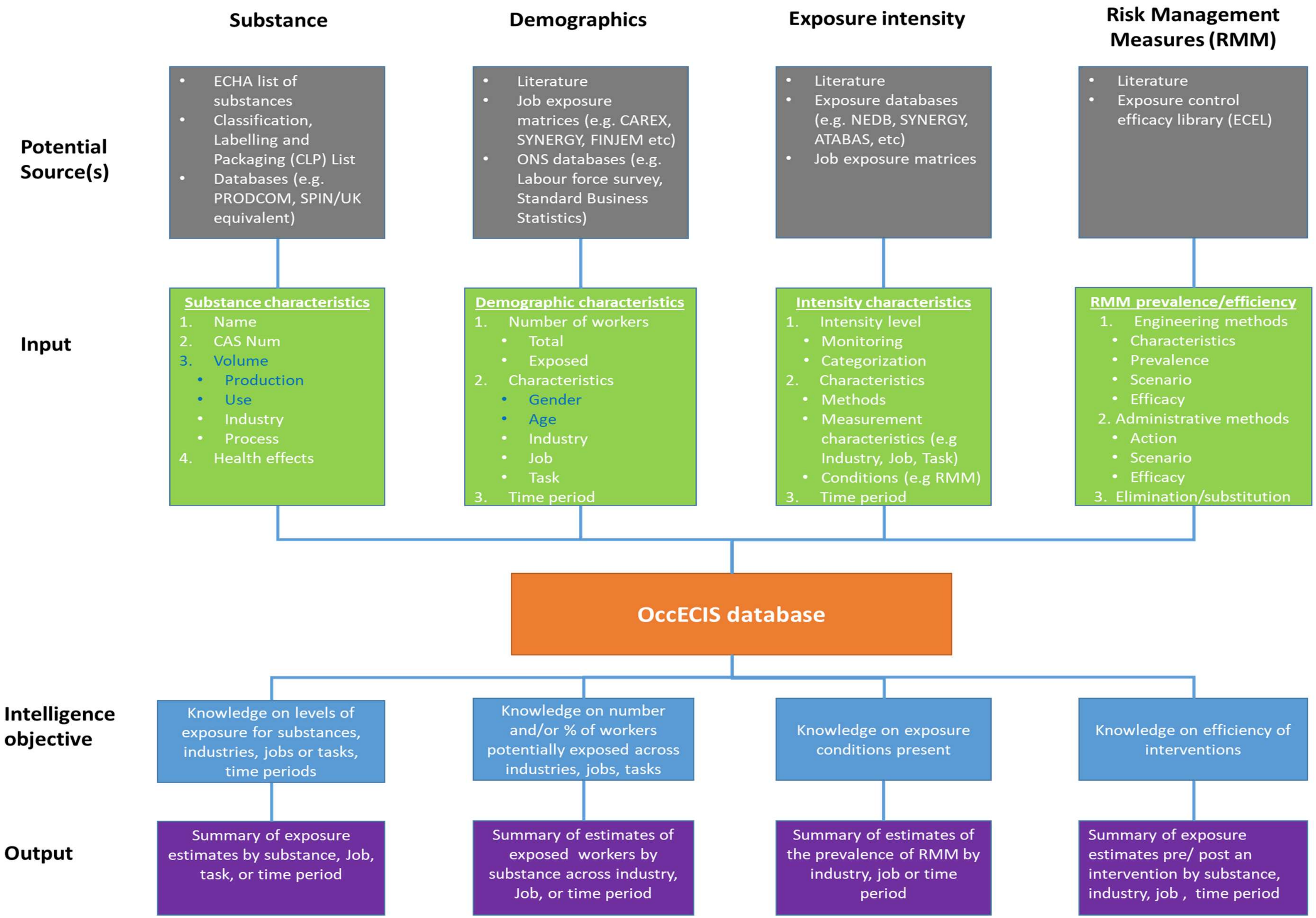
Substance
Demographics
Exposure intensity
Risk Management Measures (RMM)
Potential Source(s)
• ECHA list of substances
• Classification, Labelling and Packaging (CLP) List
• Databases (e.g. PRODCOM, SPIN/UK equivalent)
• Literature
• Job exposure matrices (e.g. CAREX, SYNERGY, FINJEM etc)
• ONS databases (e.g. Labour force survey, Standard Business Statistics)
• Literature
• Exposure databases (e.g. NEDB, SYNERGY, ATABAS, etc)
• Job exposure matrices
• Literature
• Exposure control efficacy library (ECEL)
Input
Substance characteristics
1. Name
2. CAS Num
3. Volume
• Production
• Use
• Industry
• Process
4. Health effects
Demographic characteristics
1. Number of workers
• Total
• Exposed
2. Characteristics
• Gender
• Age
• Industry
• Job
• Task
3. Time period
Intensity characteristics
1. Intensity level
• Monitoring
• Categorization
2. Characteristics
• Methods
• Measurement characteristics (e.g Industry, Job, Task)
• Conditions (e.g RMM)
3. Time period
RMM prevalence/efficiency
1. Engineering methods
• Characteristics
• Prevalence
• Scenario
• Efficacy
2. Administrative methods
• Action
• Scenario
• Efficacy
3. Elimination/substitution
OccECIS database
Intelligence objective
Knowledge on levels of exposure for substances, industries, jobs or tasks, time periods
Knowledge on number and/or % of workers potentially exposed across industries, jobs, tasks
Knowledge on exposure conditions present
Knowledge on efficiency of interventions
Output
Summary of exposure estimates by substance, Job, task, or time period
Summary of estimates of exposed workers by substance across industry, Job, or time period
Summary of estimates of the prevalence of RMM by industry, job or time period
Summary of exposure estimates pre/ post an intervention by substance, industry, job , time period

Figure 2. Conceptual framework for the development of OccECIS describing objectives and required inputs and outputs. Text in white font indicates essential attributes of the system whereas blue font indicates optional attributes.

2.5. Theoretical Questions for Outputs

The input from the EAC meetings, the funder requirements and the conceptual framework were used to formulate a series of questions that an “ideal” exposure intelligence system should be able to address. These questions were formulated on a completely theoretical basis and were grouped into the following categories:

- Exposure intensity and relevant characteristics
- Exposure prevalence and relevant characteristics
- Use (prevalence) of Risk Management Measures (type and prevalence of use) and their expected effectiveness

In addition, an ideal information system should also provide relevant background and contextual details on the hazard, the industry sector and risk management measures appropriate to the workplaces in the sector

An overview of the relevant questions for each of the above groups is provided in Table 4 below.

Table 4. Theoretical questions that an optimal exposure intelligence system should be able to answer

Category	**#**	**Question**
Exposure intensity	1	What is the distribution of exposure levels for workers of a certain industry, or occupation (or exposure scenario)?
	1a	Is the risk of exposure over and above the WEL or another threshold value beyond a certain proportion?
	1b	Has the distribution of exposure levels changed across time and including certain time-periods? If yes, how much and across which industries or occupations (or exposure scenarios)?
	1c	Based on existing trends in exposure what will the exposure levels be in the future for certain industries and occupations?
	1d	How do measured levels under a specific exposure scenario or for a specific workplace/sample of workers compare to the distribution of exposure levels for the representative group of GB workers?

Table 4. Continued: Theoretical questions that an optimal exposure intelligence system should be able to answer

Category	#	Question
Exposure Prevalence	2	What is the number, proportion (i.e. prevalence), geographical and gender distribution of workers exposed across the total GB working population and within specific industries and occupations?
	2a	Has the prevalence of exposure changed across time? If yes, how much and across which industries, or occupations (or exposure scenarios)?
	2b	Based on existing trends in exposure what will the prevalence of exposure be in the future for certain industries and occupations?
Exposure control/ RMM	3	What is the current (and historical) distribution in use of specific risk management/ control measures across industries, or occupations (or exposure scenarios)?
	3a	Can exposure be reduced in a defined scenario? If yes, what reductions can be achieved and what will the exposure levels be after?
	3b	What is the expected distribution of exposure for an industry or occupation (or exposure scenario) after the implementation of a new measure of control or a regulatory or other type of intervention?
	3c	What effect had previous policy interventions have on the levels of exposure?
Background information	4	What are the current workplace exposure limits and how have these historically changed?
	5	What are the basic characteristics of the hazardous agent, and which are the associated health effects of exposure to it?
	6	Which are the substances sources, uses and applications?
	7	In which industries, occupations and scenarios is there a potential for exposure?

2.6. Stakeholder Analysis

For an information system to be useful, it needs to cover the needs of its users and related stakeholders. We assessed the relevance of each of these theoretical questions in the previous section against the requirements for the following groups of stakeholders: policy makers, researchers, professionals (OH/H&S providers), and industry and the public. We subjectively assessed the relevance as high, medium or low, defined on the basis of potential usability of the system. The results of the assessment are shown in Table 5 (and are also available as Annex 2).

Table 5. Relevance of theoretical output questions for an optimal exposure intelligence system for different groups of stakeholders.

Question #	**Stakeholder**			
	Policy makers (incl. regulators)	**Researcher s**	**Professionals (OH/H&S providers**	**Business (industry) and public**
1	High	High	High	High
1a	High	Medium	Medium	Medium
1b	High	High	Medium	Low
1c	High	High	Medium	Low
1d	High	Medium	High	High
2	High	High	Medium	Medium
2a	High	High	Low	Low
2b	High	High	Medium	Low
3	High	High	Medium	Medium
3a	High	Medium	High	High
3b	High	Medium	Low	Low
3c	High	Medium	Low	Low
4	Low	Low	Low	High
5	Low	Medium	Medium	High
6	Low	Medium	Medium	High
7	Low	High	Medium	Medium

2.7. Data Gap Analysis

Data needs are expected to differ considerably across questions, time periods and substances. We assessed the gaps in availability of data based on the example of respirable crystalline silica (see below for details). In addition, a gap analyses for general data requirements was also performed. General data requirements were assessed in view of the questions that an optimal intelligence system should address and also the time period concerned – i.e. whether it is for estimating/assessing exposure and trends in the past or for describing the "current" exposure conditions in workplaces in Great Britain.

In principle, the available data sources and accumulated knowledge (see Sections 2.2 and 2.3) suggest that a considerable amount of data, albeit largely historical, is available in databases as well as in the scientific literature. However, comprehensive data on the current and historic use of control measures in GB workplaces appears to be lacking. Hence, a targeted data collection exercise will probably need to be carried out to collect representative data on the use and effectiveness of RMMs. Approaches for collecting the above-required data could be as follows:

- Periodic surveys across specific industries and supply chains and/or through members of important stakeholders – e.g. BOHS – or as part of ongoing initiatives such as the Exposure control indicators (ECIs) developed and utilised by the HSE as part of a number of compliance campaigns in recent years.
- Automatic collection of data through a feedback request information process as happens with COMED and/or COSHH essentials
- Through expert input from BOHS members or industrial body members who may hold such data or have good overview.
- Combinations of the above as to fill in previous and future data needs (e.g. survey or auto-collection of data for covering future needs and stakeholder members for current and previous periods).

Information on exposure levels for substances that are not process generated across certain scenarios could also be generated through the use of exposure tools and models as for example the ART (Advance REACH tool)(20).

It is important to highlight that expert input on some data aspects will be required no matter the substance, period etc. (see also Section 4). For example, the identification of exposed groups (i.e. jobs, processes, sectors) to a certain substance could be done by an already existing method, if available (e.g. a JEM,). However, this will not always be the case and for many substances such information will not be readily available (e.g. in JEMs). In other cases a method may be already established e.g. in the form of JEM but one that may be less relevant to the GB (e.g. a different occupation classification system may have been used) in which case a translation or crosswalk will be needed. The working environment and workforce are also changing, and hence periodic re-evaluations will be required in order to update any exposure rating and classifications. These essentially pertain to the need of data curators and experts (e.g. exposure scientists) to be in place in order to maintain the system but also to develop, rate and evaluate the required data whenever required.

3. Technical Development and Feasibility

3.1. Methods for Identifying Technical Solutions

We identify the core, idealized goal of the OccECIS system as answering the set of theoretical questions identified in Section 2.5. These questions serve as the basis of our technical feasibility assessment and underlying methodology of exploring possible technical solutions for the construction of the system. These considerations were the subject of joint discussions between occupational exposure domain experts and information extraction experts at the University of Manchester, with input from the EAC and HSE's Science Division.

The process included the following steps:

a) A review and analysis of previous or existing intelligence systems was initially performed

b) The results of this analysis were then used to establish the conceptual framework for the development of the system.

c) The established conceptual framework was then translated into a series of theoretical questions that the "ideal" exposure intelligence system should be able to answer, which essentially described the most optimal outputs that the system should potentially achieve.

d) These established questions were then analysed to develop mocked output web interfaces.

e) The mocked output interfaces were then analysed against the identified data sources to be optimised in terms of content and assess their feasibility.

In a separate but closely related exercise, empty MS Excel templates for data entry and storage were drafted. The structure and contents of existing large international databases such as the SYNERGY, and CODUST were used as a working example. Once developed the template contents were used to establish relational diagrams of the underlying database for the system.

Regarding feasibility, we split the technical aspects of the OccECIS system into three categories: data extraction, data storage and intelligent modelling. For each, we mention the criteria we consider in the feasibility discussion and recommendation of technological strategies (Table 6).

Table 6. Feasibility criteria across technical aspects of the OccECIS system

Technological Endeavour	**Feasibility Criteria**
Data Extraction	Availability, Structure, Access
Data Storage	Data Volume, Data Structure, Modelling Requirements
Modelling	Completeness of Inputs, Uncertainty Tolerance, Data Quality, Model Storage, Expert Constraints

In the subsequent sections, we qualify a feasibility assessment with technological suggestions for a variety of increasingly challenging cases.

3.2. Data Extraction: Requirements and Feasibility

The questions of availability of and access to relevant data sources are discussed in the data gap analysis in Section 2, and partly in the discussion of incentive structures in Section 3. The remaining technical consideration is the structure of the data sources and the data therein. All of the data source structures described in Table 7are considered feasible but increase in their cost as the difficulty of automated extraction increases.

Table 7. Assessment of the difficulty, requirements and suggested technical solutions across different types of data source structures relevant for OccECIS

Data Source Structure	**Difficulty (1-5)**	**Requirements**	**Suggested Technical Solution**
Structured Database	1	- Formal database access agreement	- Scheduled running of ETL scripts (extract, transform, load: relatively easy to outsource)
Searchable Web Source (e.g. CAREX, COMED)	2	- Negotiating direct back-end database access - Permitted automated web scraping	- Ideally, same as structured database access - Alternatively, automated web scraping using frameworks such as ScraPY
PDF Reports, Mostly Consistent Structure:	4	- Expert development of automated information extraction scripts	- Expert-crafted automated information retrieval systems - Document-based neural models such as LayoutLM, which require large amounts of training/fine-tuning data
PDF Reports, Extremely Variable Structure:	5	- Manual extraction and organization of report data	

Depending on the number and variety of data sources, building and maintaining OccECIS would most likely require a strong data curator role, which would entail activities such as:

- Mapping existing external data, critical data gaps (informed by the modeller), data licensing constraints and data integration efforts (managing updates).
- Providing technical and quality control during the data extraction and pooling exercise including by evaluating and/or updating procedures for data pooling, developing or updated variable crosswalks used for standardization, and ensuring complete removal of duplicate records when secondary data sources are used (e.g. CODUST, SYNERGY etc.)
- Creating new partnerships between academia, industry and regulators.

In terms of supporting data integration, we would recommend a lightweight pipeline, which balances a human data scientist, augmented by a simple data integration pipeline. We do not consider feasible the design of a universal data integration pipeline.

3.3. Data Storage: Requirements and Feasibility

A decision needs to be made as to the type and location of the central OccECIS database. Given the proposed scale of the system, we recommend cloud-based data storage solutions. The cost would be highly dependent on the volume, frequency and format of incoming data.

The structure of the database (e.g. Relational, noSQL) will depend highly on the structure of the data sources and their contents. For a more straightforward solution based on integrating highly structured (table-based) data sources, a relational database is sufficient. Created as part of the case study, we include an example entity relationship diagram central to a relational database that incorporates some of the example data sources we have investigated (Figure 3).

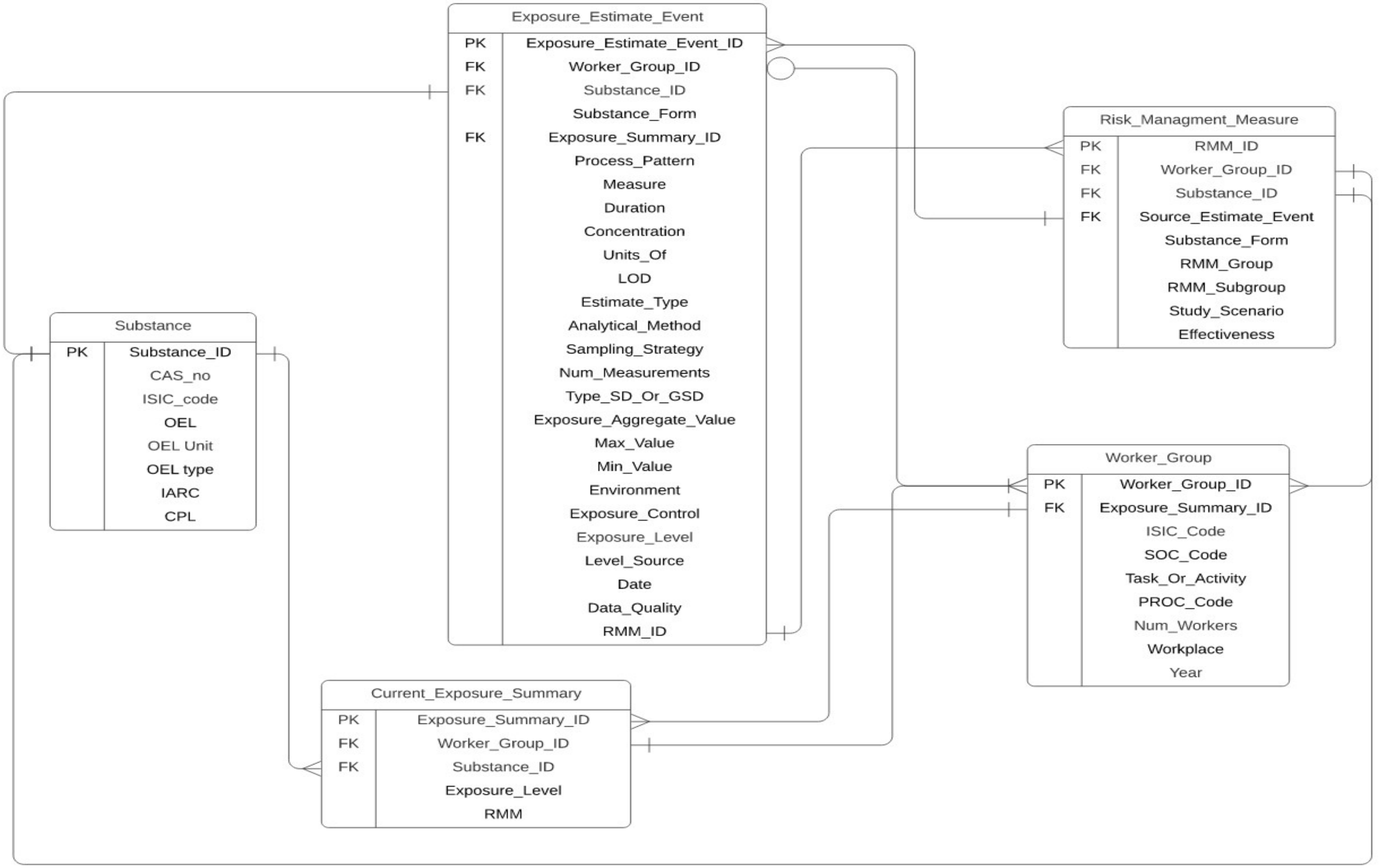

Figure 3. Example diagram describing the potential relational database structure of OccECIS

As the broader vision for OccECIS includes unstructured and semi-structured data such as text from written reports and academic literature, we recognise that a relational database structure may be too restrictive. As such, we suggest the ELK Stack as the base set of tools and services to realize OccECIS.

The ELK Stack combines three tools (Elasticsearch, Logstash, and Kibana) and provides a contemporary solution to manage both structured and unstructured datasets (covered by Elasticsearch), allowing for both structured queries and textual search. This would allow for the construction of a platform, which can target the construction of reports and visualisation over the structured data but also allow for more flexible exploration over textual data.

The Kibana component, designed to be easily integrated with the Elasticsearch backend, supports rich data visualisation, allowing for the construction of data analysis dashboards. The framework is open source, highly mature for use in industrial settings and has accessible software development expertise in the market. The close integration between data visualisation dashboards and the backend allows for more agile prototyping and reduces maintainability risks. These components are already available and are easy to deploy in cloud-based platforms.

3.4. Modelling Occupational Exposure: Requirements and Feasibility

There are three forms of modelling that can be employed to answer the target questions for the system: simple frequentist statistical modelling on top of database query responses (such as linear mixed effects regression), expert statistical modelling (expert-crafted Bayesian reasoning approaches or tools such as ExpoStats) or supervised machine learning models.

The nature and effect of the problem is such, however, that the domain experts have strongly cautioned against purely mathematical estimates that may misinterpret data. There has been strong emphasis on including uncertainty measures, data quality estimation and expert-in-the-loop assessments of all source data and modelling strategies. This is a strong limitation for machine learning approaches, which are black-box style algorithms that are difficult to interpret and cannot generally yield to expert intervention.

Hence, where estimates and predictions are required, there would be an emphasis on expert-crafted Bayesian-style models tailored to specific target outputs. This introduces a great expertise cost but may be the only solution for difficult and high-impact questions such as diagnosing the effects of interventions and predictive queries. However, it is feasible: expert tools such as ExpoStats (21) are relevant to some of the target questions, and any other models developed specifically for the

OccECIS purpose may be served in the form of an Application Programme Interface (API) which can be integrated into the system.

Suggested tools for Bayesian modelling & probabilistic programming include:

- ExpoStats (21): https://expostats.ca/site/index.html
- R (22): https://mc-stan.org/users/interfaces/rstan
- Python (23): https://pyro.ai/

Data quantity is straightforward to keep track of as one of the outputs of a given query, but data quality assessments would need to be provided at integration time with a principled and consistent expert assessment. As long as all source data is coupled with a principled quality assessment such as a categorical score (high, medium, low), any query outputs or predictions may be provided with a summary quality of the source data.

Uncertainty modelling is important to incorporate and will either be taken into account as a data input to be summarised (e.g. measurement uncertainties provided with data), as a component of simple frequentist models (summary quantities such as interquartile ranges, standard deviation) or as a part of the expert modelling problem for more sophisticated questions.

Data sparsity is likely to be a hurdle for OccECIS. The decision of how much data is enough for the modelling strategies to be reliable is in the scope of the expert modelling problems, and it may be better for the system to return “not enough data” for queries for which return a data quantity that is below some expert-determined threshold. Criteria and methods for assessing uncertainty will need be developed based on parameters such as the volume and quality of the available data and their representativeness for the scenario and population at hand. The results can then be made available to the users through the systems outputs to further support conclusions and decision-making.

In summary, we refer back to the target questions and determine whether the required modelling is within the scope of tools such as Expostats (21), simple modelling strategies or requires further expert modelling (Table 8).

Table 8. Tools and modelling strategies relevant for responding to the theoretical questions underpinning the basic framework of OccECIS.

	Modelling Requirements
1) What is the distribution of exposure levels for workers of a certain industry, occupations or exposure scenario (i.e. a specific set of conditions related to specific processes performed under certain operational conditions and in presence of specific risk management and exposure control measures)?	DB Query
a. Is the risk of exposure to silica levels over and above the OEL beyond a certain proportion?	DB Query, ExpoStats
b. Has the distribution of exposure levels for silica changed across time and including certain time-periods? If yes, how much has this change been across certain industries, occupations or exposure scenarios?	DB Query, ExpoStats
c. Based on existing trends in exposure, what will the exposure levels of silica be in the future for certain industries and occupations?	Regression-Style Trend Lines or Expert Modelling.
d. How do measured levels of RCS exposures under a specific exposure scenario or for a specific workplace/sample of workers compare to the distribution of exposure levels for the national representative group of workers?	DB Query
2) What is the number, proportion (i.e. prevalence), geographical and gender distribution of workers exposed to silica across the total GB working population and within specific industries and occupations?	DB Query
a. Has the prevalence of exposure to silica changed across time? If yes, how much and across which industries, or occupations (or exposure scenarios)?	DB Queries and Regression-Style Trend Lines or Expert Modelling.
b. Based on existing trends in exposure, what will the prevalence of silica exposure be in the future for certain industries and occupations?	Regression-Style Trend Lines or Expert Modelling.

Table 8. Continued: Tools and modelling strategies relevant for responding to the theoretical questions underpinning the basic framework of OccECIS

	Modelling Requirements
3) What is the current (and historical) distribution in use of specific risk management/ control measures across industries, or occupations (or exposure scenarios)?	(Data gap currently too severe)
a. Is there a way for reducing exposure under a defined scenario? If yes, how large reductions can be achieved and how will the exposure levels look after?	(Data currently gap too severe)
b. What will the distribution of exposure level/s be for an industry or occupation (or exposure scenario) after the implementation of a new measure of control or after a change in an existing measure of control at a national level?	(Data gap currently too severe)
c. What effect had previous policy interventions on the levels of exposure?	(Data gap currently too severe)

In summary, we recommend that in any situation where conservative (frequentist-based) statistical inference may not be a feasible framework for modelling occupational exposure, Bayesian modelling can be used as a framework, which integrates expert prior knowledge, uncertainty modelling and transfer learning from existing datasets. While each step introduces a level of extrapolation and bias, each provides a principled and realistic approach for modelling and understanding exposure on a data sparse scenario.

3.5. Incentive Structure

In the items below, we elaborate the human aspects behind the platform. These aspects complement the technical discussion and are a key enabler for the sustainability of a platform, which is highly dependent on external data. These items were collated from the discussion with the EAC and from evidence from the literature.

Instrumental to the platform is the creation of a community of stakeholders, which can support it with data and domain expertise. In order to engage and maintain a dialogue with these stakeholders, the role of a data curator needs to be formalised within the project. Commonly underappreciated as a formal role, a data curator allows the sustainability of platforms which require close dialogue with domain experts and the integration of third-party data under well-defined quality requirements. This is confirmed and it is a consensus across projects which require a community of data contributors (24, 25), and proved to be the central sustainability and growth factor of these platforms.

Identifying the set of incentives for the stakeholders is also central. Two common patterns in projects which require volunteer data contributions include recognition and purpose (24, 25). Recognition mechanisms include surfacing and acknowledging the data contributors within the platform, providing PR visibility and establishing a selective expert-level role.

The amplification of value provided by the integration between datasets can serve as a driver of purpose for stakeholders as well as the delivery of new analytical insights, which are not accessible to the data providers. Allowing stakeholders to contribute data with a minimum effort is also critical. This defines the second instrumental role within the platform (the data scientist), who can technically deliver both data integration and modelling. While this role can be facilitated with the support of a technical pipeline, we do not consider the design of a highly automated and universal pipeline feasible. Technical insights delivered by specific data contributions can be organised into reports, which can be part of the incentives structure.

It is fundamental that data contributors feel comfortable sharing data without any associated risk of reputation damage or negative exposure. It is recommended that the platform is managed by a third party, which ensures independence and anonymity when applicable.

4. Case Study: Crystalline Silica

Silica is a common mineral found in the earth's crust. Materials like sand, stone, concrete and mortar contain silica, which is released in crystalline form as a by-product during their processing. Exposure to airborne crystalline silica and particularly its respirable fraction (i.e. Respirable Crystalline Silica or RCS) is prevalent in industries such as construction. Silicates are also used to make products such as glass, pottery, ceramics, bricks and artificial stone. Inhalation of RCS can cause multiple diseases including silicosis, lung cancer, chronic obstructive pulmonary disease (COPD) and kidney disease. It is also related to the development of autoimmune disorders and cardiovascular impairment.

Previously, it has been estimated that approximately half a million GB workers are exposed to RCS during work (26). The burden of occupational cancer in Great Britain project has estimated that approximately 800 deaths and 900 new cases of lung cancer occur annually due to occupational exposure to RCS (27). Estimates of the proportion of total COPD cases or deaths where occupational exposures have contributed are uncertain and vary across a wide range of epidemiological studies. A number of reviews have estimated values of around 15%, equivalent to about 4000 deaths per year in GB (28).

4.1. Review of existing data sources relevant for RCS

Databases relevant for RCS were identified with assistance from the EAC including shared intelligence with our colleagues from Massey University, New Zealand. The databases were categorised into those providing information on workforce demographics (n=4), substance characteristics (n=5), those that were exposure databases either specific to RCS (n=3) or covered multiple substances (n=6), different intelligent systems (n=7) and JEMs (n=6). A list of all relevant data sources intensified for each of the above categories is provided in Table 9. Further details can be found on the Annex (Annex 3).

Table 9. Outline of identified data sources relevant for silica.

Name of data source	Attributes	GB data coverage	Reference
Workforce demographics / population characteristics			
Labour force survey (LFS)	Number of workers by International Standard Industrial Classification (ISIC), and Standard Occupational Classification (SOC) coding	Yes	(29)
Annual population survey (APS)	Number of workers by ISIC, SOC	Yes	(30)
UK census 2011	Number of workers by ISIC, SOC (high resolution)	Yes	((31)
Structural Business Statistics (SBS)	Number of workers by Statistical Classification of Economic Activities in the European Community (NACE) coding	Yes	(32)
Substance attributes (Background info)			
ECHA list of substances (REACH dossiers)	Substance attributes, Cas_Num, industrial sectors	NR	(33)
The Classification, Labelling and Packaging (CLP) inventory	Substance attributes, Cas_Num, hazardous properties	NR	(34)
HSE Work Exposure Limit (WEL) list	Exposure legal limit values	NR	(35)
EU exposure limit values	Exposure legal limit values	NR	(36)
US OSHA Permissible Exposure Limits (PEL) list	Exposure legal limit values	NR	(37)

Table 9. Continued: Outline of identified data sources relevant for silica

Name of data source	**Attributes**	**GB data coverage**	**Reference**
International Agency for Research on Cancer (IARC) classification list	Carcinogenic attributes classification	NR	(38)
Exposure databases - Non substance specific			
NEDB (year –)	Individual exposure measurements and contextual information	Yes	(14)
MEGA (1972-2021)	Same as above	No	(39)
COLCHIC (year –)	Same as above	No	(40)
SIREP (1996 -2005)	Same as above	No	(5)
ATABAS (1970 - unknown)	Same as above, coding by NACE and ISCO-88	No	(41)
EXPOSYN (1951 - 2012)	Same as above, coding by NACE and ISCO-89, combines data from other databases	Yes	(15)
Exposure databases - Substance specific			
ACGIHCC (construction only)	Individual exposure measurements and contextual information	No	(42)
IMA-DMP	Same as above, coding by NACE and ISCO-88	Yes	(43)
RCS in construction	Aggregate exposure measurement results and contextual data	Yes	(44)

Table 9. Continued: Outline of identified data sources relevant for silica

Name of data source	Attributes	GB data coverage	Reference
Intelligence systems			
CAREX EU	Exposure level predictions, measurement results, estimates of prevalence by NACE	Yes	(3)
CAREX CANADA	Substance characteristics, exposure level, estimates of prevalence by industry and job title	No	(4)
SIREP	Exposure levels, estimates of prevalence by industry, substance info, exposure characteristics	No	(5)
CPWR's Exposure Control Database	Exposure distributions by process, incl. a range of specific determinants. Covers silica, welding fumes. Noise and lead	NR	(7)
Silica Control ToolTM	Risk assessment tool provides distribution for exposure, summarises RMM measures and gives advise	NR	(8)
ECEL (2000~2020)	RMM efficiency by method, substance form, substance, scenario involved	NR	(6)
COMED (>2015)	Same as above	NR	(9)

Table 9. Continued: Outline of identified data sources relevant for silica

Name of data source	Attributes	GB data coverage	Reference
Job exposure matrices			
FINJEM (1945-2012)	Quantitative JEM with prediction for ISCO-68	No	(11)
SYN-JEM (1951–2010)	Quantitative JEM with prediction for ISCO-68 and region/country	Yes	(45)
DOM-JEM	Semi- Quantitative JEM with prediction for ISCO-68	No	(46)
Matgene	Exposure prediction for ISCO 1968 and PCS 1994 for occupations, and NAF 2000 for activities.	No	(47)
RCS in construction, CANADA	Exposure prediction for ISCO 1968 and PCS 1994 for occupations, and NAF 2000 for activities.	Yes	(48)
INTEROCC	Quantitative JEM with prediction for ISCO-68	No	(13)

Of the data sources listed in Table 8 some were published on the public domain or already available to the research group by being members of the development team whereas HSE has kindly provided access to work that they have previously performed. A discussion on the potential accessibility of the remaining data sources has been provided in Section 2.3.

Already available data sources that could be utilised when testing the feasibility of the system included the following:

- Labour force survey (LFS): a survey of households living at private addresses in the UK performed quarterly with the aim to provide information on the UK labour market (29).
- UK Census 2011: a decennial survey of the characteristics of the complete UK population (31).
- Structural Business Statistics (SBS): data related to the structure, conduct and performance of economic activities, down to the most detailed activity level as provided by EU Member States (including the UK) on the basis of a legal obligation from 1995 until 2018 (for the UK) (32).
- The ECHA list of substances and Classification, Labelling and Packaging (CLP) inventory: The list of registered substances contains data from the registration dossiers submitted to the European Chemicals Agency (ECHA), including information on several substance-related data, such as hazardous properties according to the classification and labelling of substances in accordance with the CLP Regulation and their safe use (33, 34).
- The International Agency for Research on Cancer (IARC) classification list: A list of substances, mixtures, and exposure circumstances classified by IARC monographs on the basis of their carcinogenicity and the classification assigned (38).
- The HSE Workplace Exposure Limit (WEL) list: A list of the most recent British occupational exposure limits set and used with the Control of Substances Hazardous to Health Regulations 2002 in order to help protect the health of workers in GB (35).
- “RCS in construction” database: a literature based exposure database of RCS levels in the construction industry containing 6118 records (2858 of respirable crystalline silica) extracted from 115 sources, summarizing 11,845 measurements collected between 1978 and 2010 (44).
- An extract from NEDB (National Exposure DataBase) containing approximately 90 measurement collected during work in brick manufacturing (14).
- The INTEROCC JEM: A quantitative job exposure matrix with a period prevalence and intensity axis developed on the basis of the Finish Job exposure matrix (FINJEM) and coded in the International Standard Classification of Occupations 1968 (ISCO68) (13).

4.2. Tailoring the conceptual framework to RCS

On further analysis, the developed framework was tailored to the working example of silica in view of the data sources described in Section 4.1. Essentially the tailoring process involved the “fitting” of the data sources identified and summarised in Table 9 with the conceptual framework described in Section 2.4. These included the establishment of the relevant tables to be created, the mapping of the relationships between them as well as the identification of the parameters that will be used to link the different data tables. The results of this process are schematically summarised in Figure 4.

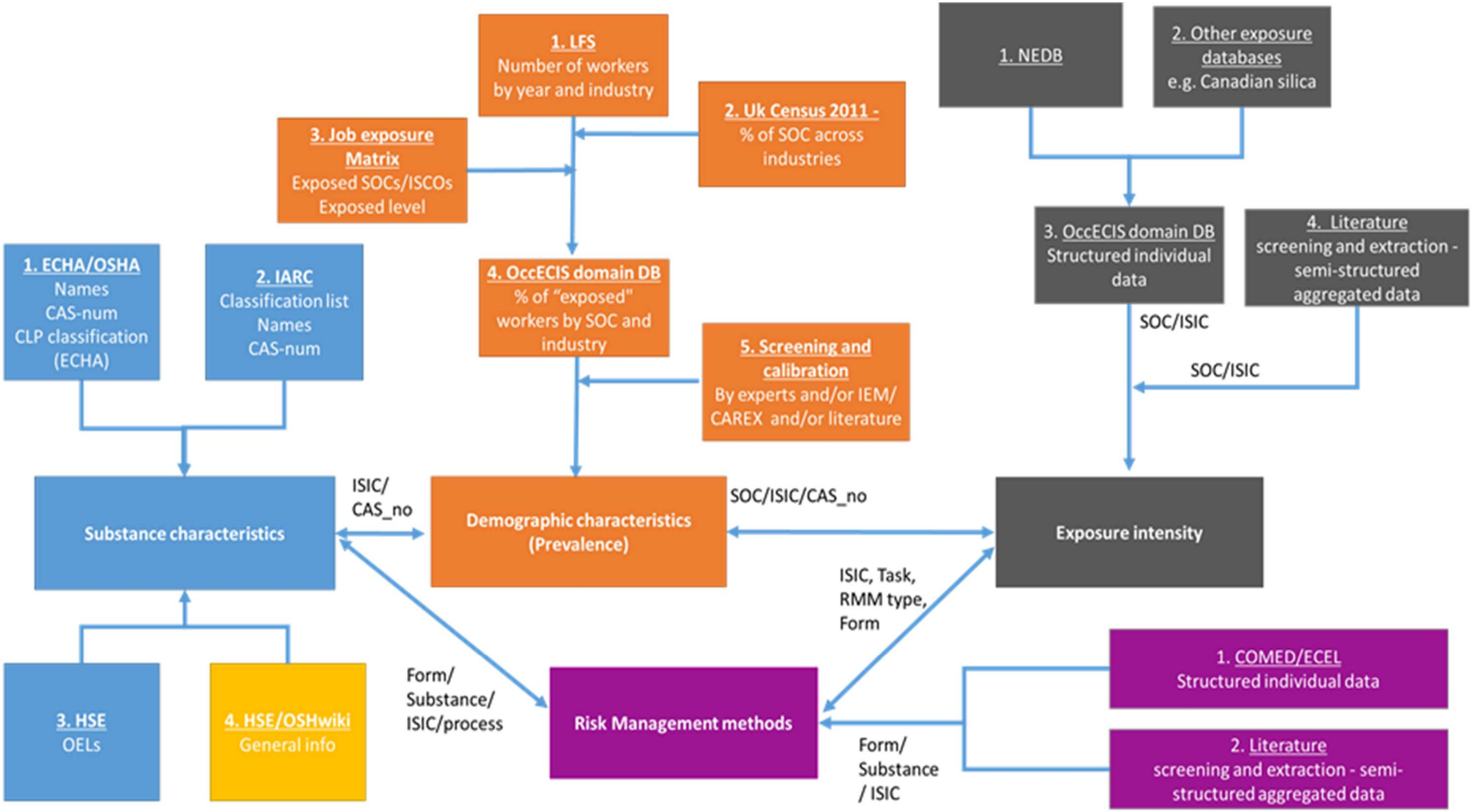
1. LFS
Number of workers by year and industry
2. Uk Census 2011 - % of SOC across industries
3. Job exposure Matrix
Exposed SOCs/ISCOs
Exposed level
4. OccECIS domain DB
% of "exposed" workers by SOC and industry
5. Screening and calibration
By experts and/or IEM/ CAREX and/or literature
1. ECHA/OSHA
Names
CAS-num
CLP classification (ECHA)
2. IARC
Classification list
Names
CAS-num
1. NEDB
2. Other exposure databases
e.g. Canadian silica
3. OccECIS domain DB
Structured individual data
4. Literature
screening and extraction - semi-structured aggregated data
SOC/ISIC
SOC/ISIC
Substance characteristics
ISIC/
CAS_no
Demographic characteristics (Prevalence)
SOC/ISIC/CAS_no
Exposure intensity
ISIC, Task, RMM type, Form
3. HSE
OELs
4. HSE/OSHwiki
General info
Form/
Substance/
ISIC/process
Risk Management methods
Form/
Substance
/ ISIC
1. COMED/ECEL
Structured individual data
2. Literature
screening and extraction - semi-structured aggregated data

Figure 4. Draft general design, information pooling and workflow of OccECIS for RCS.

4.3. Data gap analysis

A data gap analysis was performed at the macro and (partly) micro-level using the data sources described in Table 9 (Section 4.1) and the principles outlined in Section 2.6. Overall, for exposure intensity, prevalence and background information, available data appear to exist to an extent that enables descriptions of temporal trends in exposure across many of the important industries and occupations for silica exposure. This includes British settings where coverage appears to be provided by the NEDB and the more recently established IMA-DMP database. It's worth mentioning that IMA-DMP contains more than 5,500 personal measurements collected from UK members of the European Industrial Minerals Association (IMA-Europe) during the period 2002 and 2016 (43). Similarly a large database summarising the literature reported measurement data on silica exposure in the construction industry has also been established by the University of Montreal – i.e. the RCS in construction database (44). Despite the above, an in-depth mapping exercise will require complete access to those and other available databases and hence some caution on the above interpretation is required.

As noted in Section 2.6, a clear gap in data exists concerning the prevalence and distribution of RMMs within workplaces. Limited data may be available on exposure control (e.g. on exposure control indicators – ECIs) from inspection reports and compliance campaigns carried out by the HSE or from internal documents / reports held by trade organisations or companies. However, it is unlikely that these data sources will provide a representative and comprehensive overview of RMMs used.

In addition, although historical data for silica appear sufficient with respect to intensity of exposure it is known that the collection of exposure data within GB has reduced considerably in recent decades. For example the NEDB was reported to contain >80,000 measurements in the early 2000's with most of those data though collected in the period between 1985 and 1990 and additions in the range of a few hundred personal exposure measurements added every year following this period (1). This pertains to the need for supporting the system with GB measurements in the future both in relation to trends in silica but also for other exposures. This can be achieved through both the performance of additional targeted campaigns (e.g. for emerging risks) by the regulators as well as by the contribution with data from private holders such as independent research institutions, the industry, and H&S and OH providers. For the latter (i.e. privately held data) data ownership and handling will likely be subject to contractual terms between the providers and their clients and, possibly also the General Data Protection Regulation[4] which may impact on the sharing of such information with the system. Similarly, the potential of measurements showing deviations from the current legislation (i.e. exceedance of established

[4] https://www.gov.uk/government/publications/guide-to-the-general-data-protection-regulation

WELs) may discourage private providers from sharing their data. Approaches including incentives to provide data would also need be developed.

The complete results of the micro-level gap analysis are available in Annex 3. The results of the macro-scale analysis were required to identify gaps across the domains defined by the conceptual questions established in the developmental process of the system framework (Section 2.4). In this line, the previously mentioned data requirements and classifications as well as their inter-relationships can be visualised in the form of a matrix, where on the y-axis you have the data domains and, on the x-axis, you have the time-periods. This matrix is visualised below in Table 10.

Table 10. Macro level gap analysis results with focus on RCS

Output question category	**Time period**	
	The past	**The "current" (and the future)**
Substance characteristics (Background information)	Basic attributes (e.g. CAS-num, Carcinogenic classifications, irritant classifications etc.) available. Detailed descriptions not available.	*Short-term (present):* good data are available Long-term (future): Periodic updates of the toxicity, carcinogenic, irritant classifications etc. will need performed.
Demographics	Good data available at industry and occupation level. Identification of exposed populations may be an issue depending on the substance involved. For silica this can take place on the basis of an exposure matrix. Prevalence of exposed workers will require in most cases data from the literature or expert assessments.	*Short-term (present):* Pretty good in relevance to the demographics. Some agreements and accessibility to ONS databases will be required. *Long-term (future):* Periodic updates of estimates of prevalence (assigned manually/semi-automatically) will be required. A mechanism to capture and integrate annually released data from the respective databases will also be required. Bayesian analytical approaches for trend estimations will also be required.

Table 10. Continued: Macro level gap analysis results with focus on RCS

Output question category	Time period	
	The past	The "current" (and the future)
Exposure intensity	For silica there are relatively good historical data for several industry/ occupation/scenario combinations. A detailed analysis will be required to see the coverage of this. For other substances availability will vary considerably. Detailed data mappings will need be performed on an ad-hoc basis. Literature reviews will need be performed whenever individual data not available. A mechanism for read across can be developed based on substance forms, and scenario similarities. This, under conditions, could be standardised and "automatically" applied.	*Short-term (present):* Situation is unclear, but for some industries beyond construction relevant data are available (e.g. NEPSI/ IMA-DMP) *Long-term (future):* A mechanism and investment for ensuring the capturing and consistent update of the (national and other) databases will be required. Agreements of data sharing with external stakeholders holding relevant data will be required. Bayesian analytical approaches for trend estimations will also be required.
Risk Management Measures	*Prevalence:* No data are available or, at least, accessible, regarding the historical prevalence of RMM in GB workplaces. *Efficiency:* Two intelligence systems have been developed and are currently available that provide information related to these: COMED and ECEL. We have been discussing collaboration with COMED which is more dynamic and complete than ECEL.	*Prevalence:* Collection of information regarding the prevalence of RMM in GB workplaces will be required regarding the current and future time periods. This could be done e.g. by building on HSE ongoing initiatives (e.g. ECIs), with the performance of periodical surveys or the collection of information in a self-reporting scheme as part of receiving feedback from the system. *Efficiency:* The intelligence systems available can serve the required purpose. Support of the connected systems to perform any required future updates may be needed.

4.4. Testing the system’s feasibility

We carried out a pilot exercise to implement the workflow and data processing described in Figure 4 (Section 4.2) to test the feasibility of the system. We used the data readily available as described in Section 4.1 focusing primarily on the prevalence and intensity of exposure to RCS within the construction and brick manufacturing industries, respectively.

We estimated the prevalence of exposure as follows:

Step1: Data on the number of workers across each industry in GB for the period between 2008 and 2019 were extracted by the Labour force survey (LFS). Since the specific data are controlled within UK data services (i.e. ONS), at least to the detail required for our analysis, these data were temporarily extracted from the EUROSTAT databases.

Step 2: Data on the distribution of workers with certain occupations across industries were extracted from the UK Census 2011.

Step 3: This census-derived distribution was used in conjunction with the LFS data to estimate the annual number of workers for each occupation within each industry for the defined time period.

Step 4: We extracted data concerning the occupations exposed to silica from the INTEROCC JEM (13). Since INTEROCC is a JEM coded in the ISCO-68 coding system, we had to translate the exposed codes to the Standard Occupational Classification (SOC) 2010 coding system used by the UK data services.

Step 5: Once INTEROCC exposure status was assigned then the number of exposed workers per occupation was estimated. For this, we used the exposure prevalence estimates for the last time period covered by INTEROCC (2001-2003). Prior to implementation a conceptual evaluation of the properness of the INTEROCC proportions of exposed workers in relevance to the occupation involved was performed by an experienced exposure scientist, member of the research team.

Step 6: The INTEROCC JEM besides estimates of the exposure prevalence also includes quantitative intensity estimates for each agent concerned. These intensity estimates were used to classify occupations as being high, moderate, or low exposed to RCS in direct comparisons with the existing WEL of 0.1 mg/m^3. In particular, the following logic was applied:

- High = Average RCS level >0.1 mg/m3
- Moderate = Average RCS level between 0.05-0.099 mg/m3
- Low = Average RCS level < 0.05 mg/m3.

This way the number of exposed workers per occupation and industry as well as their level of exposure was estimated. The above process is schematically summarised in Figure 5.

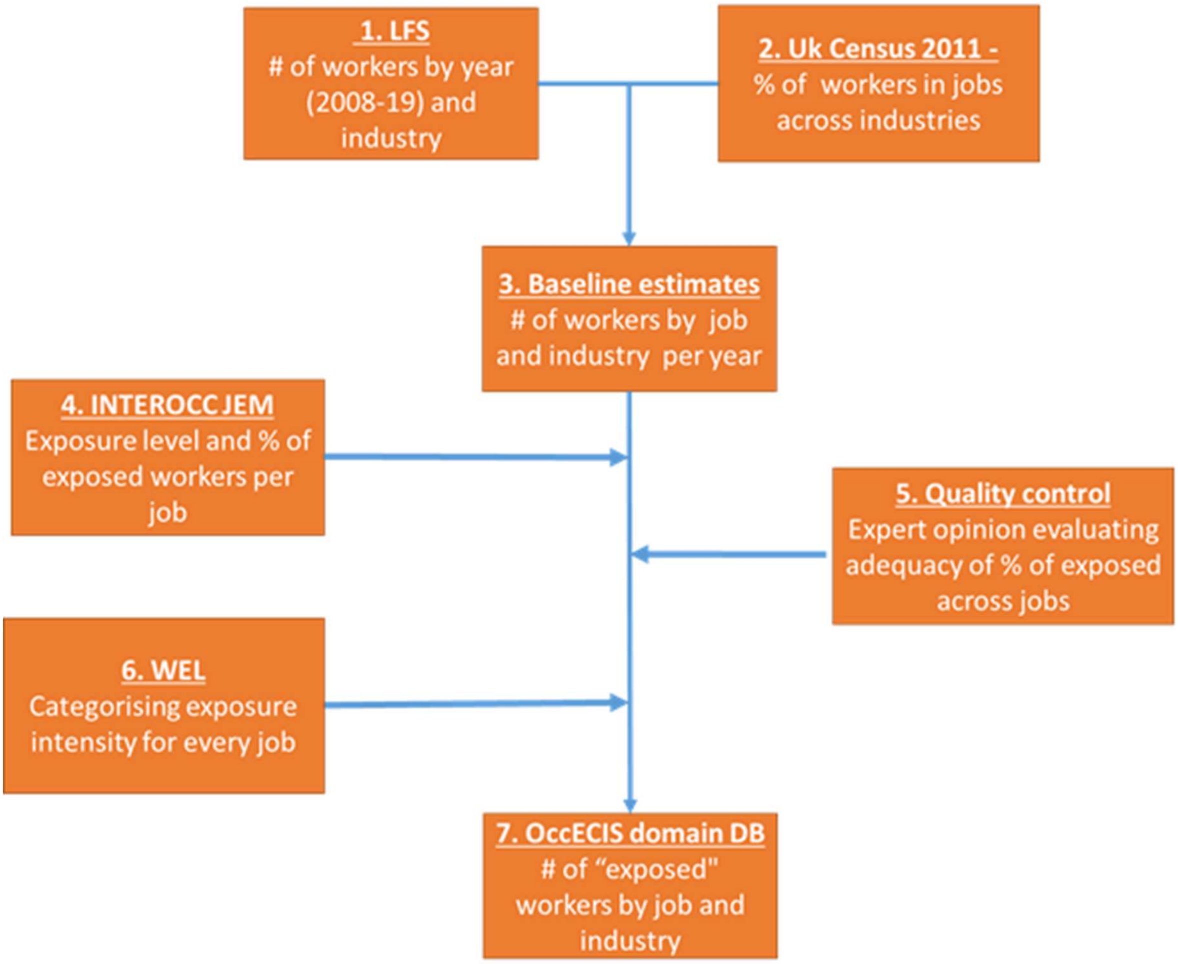
1. LFS
# of workers by year (2008-19) and industry
2. Uk Census 2011 -
% of workers in jobs across industries
3. Baseline estimates
# of workers by job and industry per year
4. INTEROCC JEM
Exposure level and % of exposed workers per job
5. Quality control
Expert opinion evaluating adequacy of % of exposed across jobs
6. WEL
Categorising exposure intensity for every job
7. OccECIS domain DB
# of "exposed" workers by job and industry

Figure 5. Examine the feasibility of estimating the number of exposed workers to respirable crystalline silica for every job involved in the construction industry (LFS = labour force survey; WEL = Workplace exposure limit).

For exposure intensity, the complete RCS in construction database (44) was available alongside an extract from the NEDB of approximately 90 measurements of RCS from the "Manufacture of bricks, tiles and construction products, in baked clay" sector were made available by HSE. These NEDB measurements covered a small number of activities during brick manufacturing and were collected in the period between 2012 and 2018. The "RCS in construction" is a literature-based exposure database covering the breadth of the collected information in the period between 1974 and 2010 regarding exposure levels to RCS in the construction industry. Standardisation and coding of all available data was a logistical issue in terms of established time frame for the delivery of the present project particularly when concerning the aggregated data contained within the RCS in construction database. We thereby proceeded in coding according to the SIC and SOC 2010 coding systems only the NEDB measurements.

The above were subsequently used to illustrate examples of relevant outputs for the system in terms of both prevalence and exposure intensity. It must be noted that the above exercises do not represent a final proposal for the estimation of the current and historical prevalence of exposure to RCS in GB. Instead, they should be seen as a simple example on the basis of the limited data accessible and manageable at present demonstrating that the development and functionality of the system is feasible. The availability of further individual data alongside a more detailed data rectification and treatment process will enable a more adequate analysis of the data to be performed. This will include estimates of the quality of the inputs, and of the uncertainty surrounding the derived exposure estimates in respect to the variability of exposure within and between workers.

Figure 6 provides the estimated number of workers exposed to RCS (according to the INTEROCC JEM) by occupation in the construction industry. Builders (i.e. construction and building trades n.e.c), labourers (elementary construction occupations), carpenters and joiners, and bricklayers and masons account for the majority of the RCS exposed workers in this sector.

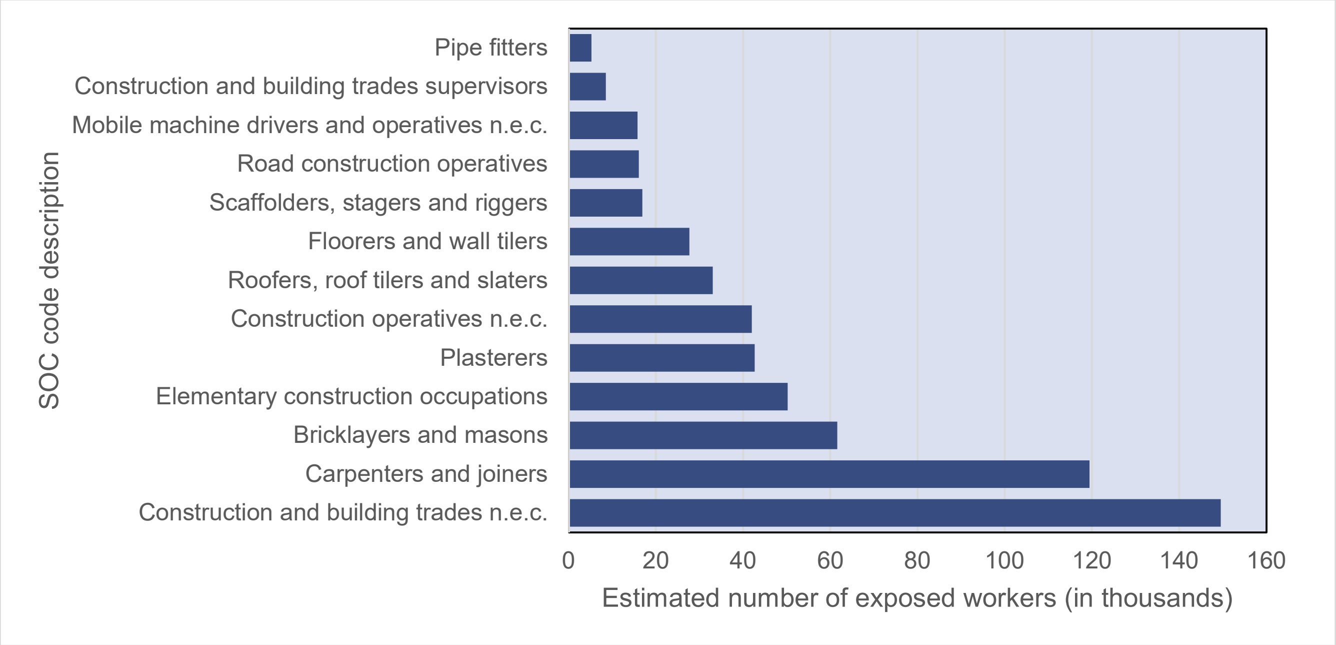

Figure 6. Estimated number of exposed workers for occupations within the construction industry for the year 2011.

Figure 7 summarises the trends in prevalence of exposure within the 5 occupations with the most exposed workers in construction. It can be observed that the prevalence of exposure within the industry reduces between 2008-2010 possibly as result of the 2008 financial crisis. Subsequently, and as the economy recovers, the numbers of employed and consequently of the number exposed workers increases at least for the three most prevalent occupational groups.

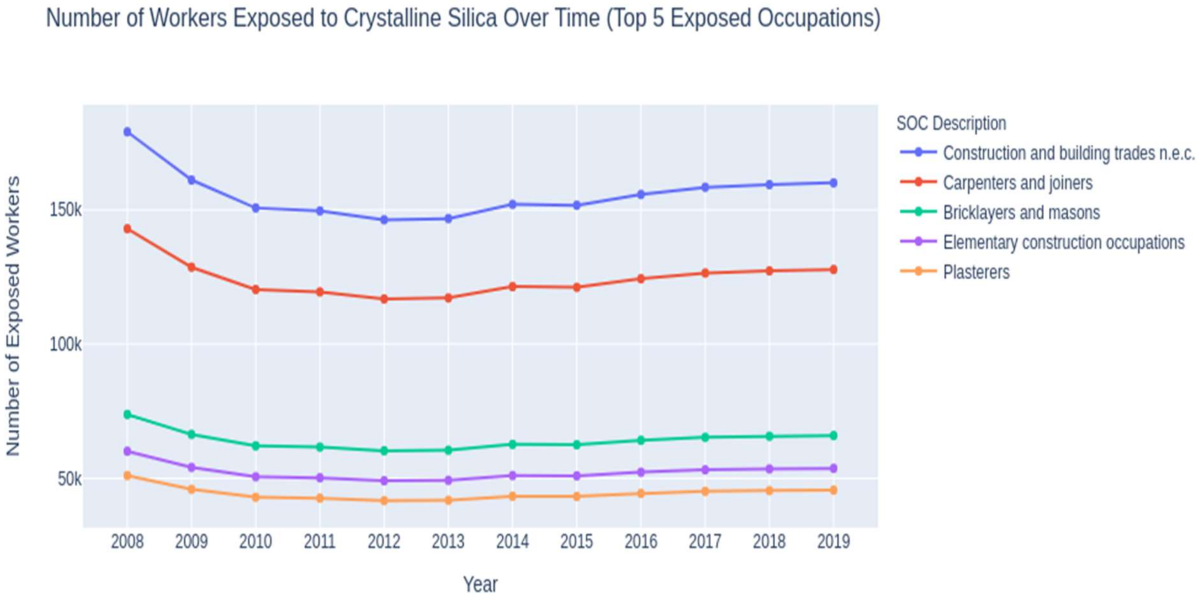

Figure 7. Estimated trends in the number of exposed workers for the five occupations with the most exposed workers within the construction industry.

A summary of the exposure levels within occupations included in the NEDB extract available is provided in Figure 8. Results of the 87 personal measurements collected in the years 2012-2018 available are shown relatively to the established WEL (red line). As it can be seen mean levels of exposure for glass and ceramic makers, decorators and finishers exceed the available WEL, whereas the same applies for a large proportion of the collected measurements among process operatives. On the contrary, forklift drivers and production managers seem to be exposed to very low levels of RCS during work. Amid the small number of measurements available, no attempt to analyse trends in exposure was undertaken even graphically. The availability of further data should be expected to further optimise data visualisations.

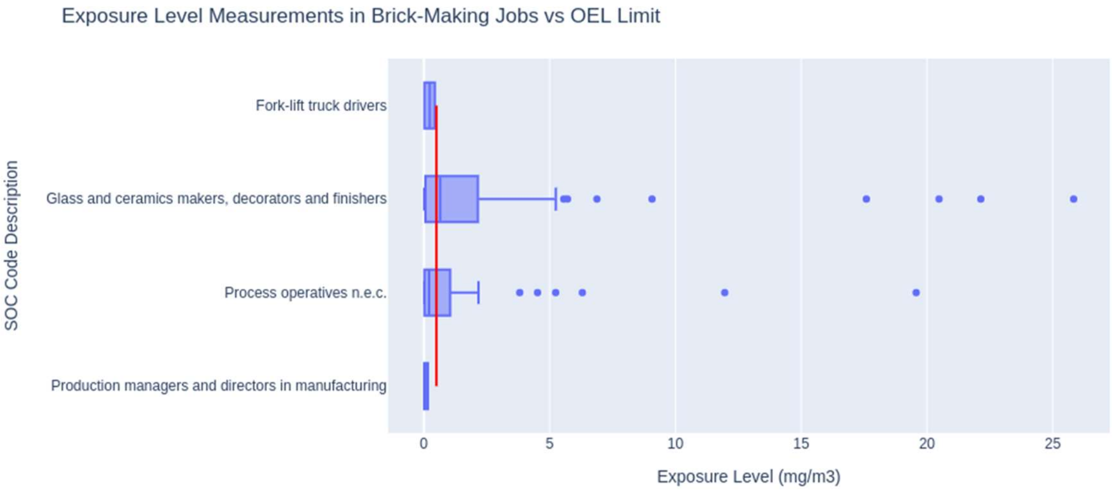

Figure 8. Exposure levels to RCS across different jobs of workers in the brick manufacturing industry. The red line represents the current WEL for RCS in GB (0.1 mg/m^3), and the dots represent values for the highest exposure measurements for that occupation.

5. Feasibility Evaluation with Respect to Original Research Questions

Within this section, we address each of these secondary objectives that were provided in Section 1 detail below.

To describe the available data on agents that cause work-related respiratory diseases

We have identified a number of data sources, which builds on previous work done for a project funded by EU-OSHA. Annex 1 provides a detailed overview of the relevant data sources, including describing population characteristics, substance attributes, non-substance-specific exposure databases, substance-specific databases, other intelligence systems and job-exposure matrices.

It is acknowledged that a large amount of data are available, although much will be of a historical nature going back 30 to 40 years and important data gaps exist for current exposure levels. We would recommend that data going back no more than about 20-25 years should be included in OccECIS.

To determine whether the required data sources are available

For silica and the other agents that cause work-related respiratory diseases, based on the gap analysis and data mapping exercises we performed (see Sections 2 and 4) we believe it is feasible to establish lists of reliable sources of data for exposure. However, these data are likely to be unevenly spread across exposure scenarios, and for some perhaps rarer exposures (e.g. food flavourings, platinum salts), significant occupational hygiene and occupational exposure science resource will be required. Data collection should ensure that data are of sufficient quantity and quality, and that data can be organised into a format compatible with OccECIS. Some data, such as those held by the industry, or research organisations may not be publicly available, or may require funding to access, though unlikely to be at prohibitive levels.

Another important consideration is that the data may be held as individual measurements or in a variety of aggregations, by some of the stratification factors mentioned in answer to the previous question. Robust procedures will need to be developed to combine individual with aggregated exposure data.

For RMMs, including both prevalence and impact in exposure levels at a population level, there is limited, if any, information available. Further research will be required to determine whether relevant data are available within private stakeholders and to develop new initiatives for data collections. A sensible starting place may be a review of the intelligence provided by HSE's ECIs. However, the support of a system like OccECIS with new exposure data collection initiatives will be an important component.

To describe data gaps on occupational exposure to substances, in terms of prevalence and intensity

Both macro (e.g. substance, industry) and micro-scale (e.g. individual occupations/ processes, periods of coverage) analyses of the data gaps are required. We have (partly) carried out such an analysis for silica, but it needs to be undertaken for other priority agents. Criteria will need to be developed to identify the priority data gaps that need to be addressed for the system to fulfil its purpose (i.e. to provide adequate responses to the theoretical questions discussed in Section 2). This will include numbers estimated to be exposed, level of exposure, feasibility of implementing different risk management measures and perhaps other factors (to be determined).

A gap analysis has been undertaken for silica, which has been included in the Annex 3. It should be noted that availability of data does not necessarily mean relevance of data to specific scenarios since e.g. periods of coverage may vary. The gap analysis also includes a stakeholder relevant analysis.

To determine the available intelligence on what risk management methods are in place in different sectors or occupations to control or reduce exposure levels

Our gap analyses suggest that the available data related to the prevalence of specific RMM in British workplaces is rather limited. This is in contrast with the efficiency of different exposure control measures where several dedicated systems and databases are available ranging from generic ones (e.g. ECEL, COMED) to systems specific to certain exposures and/or industries (e.g. Silica Control ToolTM, CPWR's Exposure Control Database). Of those, COMED has been identified as probably the most promising providing a very thorough design with an elaborate data evaluation process and an interface that enables the dynamic expansion of the system though a live data capturing mechanism. Queries through the interface that allow visualisation of the data under specific scenarios (e.g. process, sub-process, RMMs, data quality) work as the incentives for both the potential user and/or data providers. Unfortunately, only a few hundred of scenarios has than far been integrated to the system which is still under development.

Targeted data collection exercises will be required to cover this lack of intelligence regarding the presence and prevalence of RMMs in British workplaces. Potential approaches that can be utilised in this line and gather the required data are summarised within section 2.7. Any future efforts to collect these data should include elements that will allow the periodic update of the information held by the system in its database as to ensure that the evaluation of potential intervention efforts are properly supported.

To determine how different types of data can be captured most efficiently and integrated into the database

Clearly, some data capture methods are more efficient than others. We would propose that systems are developed for capturing structured and unstructured data that exists either in aggregated or summary form or as individual exposure measurement. It is important that contextual information is captured also. We believe that occupational hygiene and statistical expertise are required to appropriately capture issues of data quality, bias and uncertainty.

We recommend negotiating access to structured database APIs wherever possible, investing in the development of specialized information extraction algorithms where PDF-style data sources are sufficiently regular in structure and serving the extraction models as an API that used to integrate new data into a NoSQL database via an ELK stack. Lastly, we emphasise the strong need for a data curator role that can maintain the automated data flow and recognize when changes need to be made, or where manual extraction/input is the only option for highly unstructured new inputs. This includes the identification of exposed groups (i.e. occupation, industries), tasks and processes as well as the estimation and quality control of the estimates for the proportions of exposed workers whenever required.

To determine how to make the system dynamic and easily updatable

There probably needs to be defined minimum data standards for which any of the OccECIS proposed standard outputs can be updated using appropriate mathematical and statistical models that should be developed. However, for most situations or scenarios, it is likely that appropriate judgement will need to be made about any new data added to the system (quality, bias, uncertainty) and how it is related to existing data in the system. In some circumstances where the data are more descriptive than numerical, then qualitative approaches may need to be employed. Expert hygiene and mathematical/statistical judgment may be required.

The data curator role described previously would also interact with domain experts on suggested updates to the system and integration of new data sources. For most cases, however, data input should be encouraged through a structured front-end service (in collaboration with domain expert assessors) supported by strong incentive structures and good relations with industry role-players.

To define how exposure-control data can be analysed and exposure prevalence and intensity (high/medium/low) be determined by sector/industry, occupation, age and time period

Exposure intensity will be categorised in general terms as follows. For substances for which workplace exposure limits are available defined cut-off levels for high, medium and low exposures can be established. For example for respirable crystalline silica the cut-off points could be developed based on the UK WEL of 0.1 mg/m^3. Alternative approaches are also available, and the final approach chosen may need to be tailored to the specific substances or exposure circumstances. Some workers may experience exposures that are highly variable between and within working days and monitoring may be performed under worst-case scenario approaches (i.e. by selecting and monitoring during high exposed tasks) which may lead to results that are not representative of daily working exposures. To reduce this potential the occupational groups of interest could be assigned an exposure level following an analysis of the exposure distribution within the measurements available. The level can then be assigned on the basis of a defined proportion of the available measurements that exceed the chosen cut-off limit. Alternatively, information related to the between and within worker variation could be integrated to weigh the mean estimates and properly assign the group to an exposure category.

Such an approach of course will need to further account for the absence of measurement data and/or the presence of non-UK data on any sample (i.e. the grouping will need to be representative of the UK conditions) and expert opinion could be utilised in this case. In the absence of WELs an appropriate percentile of the exposure distribution could be utilised.

For some other substances for which exposure measurements may be inadequate to characterise exposure, or for which no thresholds can be defined (e.g. antineoplastic drugs) expert opinion combined with data on indices such as the frequency of exposure. Similar approaches can be implemented to account for exposure that may generally be low or characterised by an intermittent nature and thereby high variability in day-to-day and between workers exposures (49). Exposure intensity banding philosophies on the basis of exposure control methods applied and their adequacy may be used.

To make recommendations on how data quality should be assessed

We propose that standardised techniques be developed and used for assessing data but also output quality. In principle, for measurement data, quality can be evaluated on the basis of standard approaches accounting for limitations in the methods used to collect the data (e.g. type of sampler, analytical method applied, etc.), the sampling strategy applied, and the completeness in terms of contextual information reported. The work previously performed by Tielemans et al., (50) can form the basis for the development of a framework that can be utilised.

Information availability is inherently associated with the study aim and OccECIS is a system that will serve multiple aims. It is thereby essential that data are not excluded a priori on the basis of missing information. Instead a core set of key contextual information that need be available to define a minimum level of quality for data to remain in the database will need be established - this is essential in defining some of the future uses of the system (i.e. the required information will be different for modelling trends or for evaluating the effectiveness of interventions at a workplace population level). For implemented sampling and analytical methodologies, these will need be considered in terms of their adequacy alongside the representativeness of the results in terms of the variability present in the workplace. The latter is highly relevant particularly when sourcing aggregated data from the literature, in which case the reported estimates may be subject to bias sourcing from e.g. the size of the involved sample (i.e. variability), the method for selecting workers for measurements, the sampling duration involved etc. For those data included in the system, quality criteria will also need to be developed to enable appropriate interpretations to be associated with any analyses derived by the system.

The representativeness of the data in terms of the UK conditions will also need to be assessed. This can be achieved by the development using literature data of a ranking system (most comparable to least/non-comparable) accounting for differences in working conditions and production systems between countries. In principle, the quality of the data and outputs will gradually reduce depending on whether non-UK, or non-substance specific (i.e. read-across) methods, model estimates or expert judgement are used. Although, we expect that a combination of methods will often need to be applied to address data gaps. Data quality can be visualised in a simple relationship chart (see Figure 9).

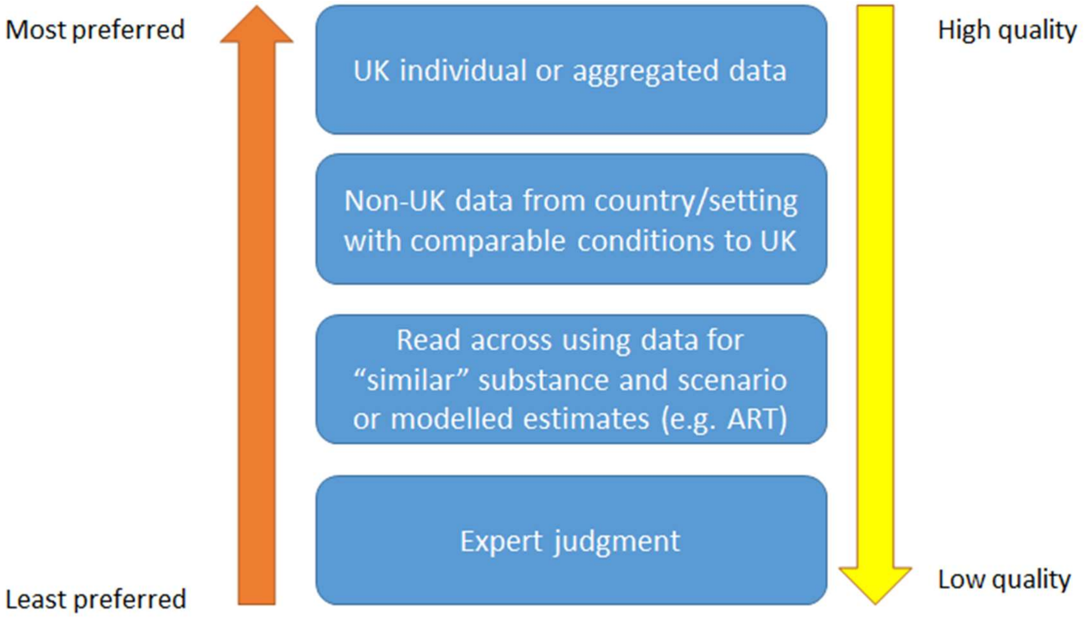
Most preferred
UK individual or aggregated data
High quality
Non-UK data from country/setting with comparable conditions to UK
Read across using data for "similar" substance and scenario or modelled estimates (e.g. ART)
Expert judgment
Least preferred
Low quality

Figure 9. Relationship between data source and quality of information attached.

To propose a methodology for how will trends over time might be assessed

Here we need distinguish between two types of trends:

1. Exposure prevalence
2. Exposure intensity

For the first of these, trends in number of workers in an occupational or industry can be carried out using data from official statistics. These included data from the Census that takes place every 10 years, in combination with the Labour Force or other relevant national surveys.

For the second, for substances for which sufficient measurement data are available, classical approaches of modelling trends could be employed. This involves the application of linear mixed effect regression analysis, and/or of general additive modelling, to examine patterns of change in exposure levels across years or well-defined time periods (intervals). The applied models account for correlations between repeated measurements from the same individual works, workplace, industry or occupation involved (i.e. the so called “random” effects) as well as differences sourcing from workplace characteristics (e.g. presence of LEV, nature of process, company size etc.) or the sampling methodologies involved (e.g. type of sampler, duration, fraction, etc…).

For substances for which insufficient measurement data will be available historical trends in exposure could be determined using expert statistical modelling approaches (expert-crafted Bayesian reasoning approaches) or by read across approaches. The latter could be based on trends estimated externally (i.e. the literature) for the specific substance in question (e.g. RCS in the minerals sector; Zilaout et al., 2020 (51) or for general trends in inhalation exposure (e.g. (52). Further work may need to be carried out to assess if these trends have persisted. If needed similar reviews as the latter can be performed to update the existing evidence to the present situation.

Associations related to the efficiency or effectiveness of interventions can be analysed using standardised statistical approaches, which will depend on the nature of any intervention. To date, this has been the use of linear-mixed effect regression models for pre and post intervention approaches for estimating efficiencies on a sample level and then extrapolation of the results on the general workplace population of the GB. Other approaches might also be available and further work would be required to characterise these.

To determine plausible uncertainty ranges for exposure-control prevalence and exposure-control trend estimates

Although some data in relation to the efficiency and surrounding uncertainties of different control measures is available in databases such as COMED, there is little information about their actual prevalence and working status/efficiency in the field. Given the likely lack of data on risk management measures, this question is probably the one that is hardest to respond to and for which feasibility is likely to be at its most problematic. Approaches may need to be developed on the use of RMMs and their general effectiveness. This could be in the form of worker surveys such as through OccIDEAS (53), through company surveys (in line to the previously mentioned HSE ECIS initiative or similar to the spot checks currently carried out by HSE in relation to RMMs to control COVID-19) or through expert elicitation (e.g. the BOHS membership).

6. Developing the platform: estimated effort

Considering this feasibility analysis, it is possible to estimate the initial effort involved in the construction of an initial or prototype version of OccECIS. We estimate that a fully functional version of the platform could be developed in an 18-month project. As reflected in the analysis, a domain expert serving as data curator (along the duration of the project), a software developer (8 months) specialised in the ELK stack and a data scientist (10 months) could provide a fully functional version of the platform.

We propose an iterative method where an initial version of the system could be developed for a specific subdomain (respirable crystalline silica), having a functional end-to-end demonstrator within the first 6 months. This would allow for the collection of expert feedback at an early stage. We would also recommend that such a project have an advisory committee with representatives from BOHS, academia and industry, who would be engaged as data stakeholders and would support in the co-design of the platform, thus making the end product acceptable to potential data contributors and users of the systems outputs.

This report provided a concrete de-risking of the main design questions behind the platform. This would allow for a software development phase with a clear set of requirements and with a well-articulated high-level specification.

7. Discussion, Conclusions and Recommendations

We believe that, in principle, it is feasible to develop a system that can be used to monitor the effectiveness of any policy intervention. Although, it is clear that the quality of input data will largely determine the quality of the outputs of the system. Note that alternative arrangements will need to be put in place to monitor other aspects of any intervention, such as psychosocial aspects.

It is important that such a system is not just a repository for data, but that it can be populated with data such that it provides added benefit to simply storing data as currently happens with databases such as the National Exposure DataBase (NEDB).

If a decision is made to develop a prototype of the system, it would be important to identify priority agents in addition to RCS could be identified in order to start developing and populating the system. Alternatively, building a prototype with just RCS could be an option. Data such as lists of scenarios with associated demographic data and information on the health effects of the agents might be usefully used to populate the system initially.

The next step would be to identify and capture all available data for those agents that HSE have identified, such as silica, as being of highest priority (or, alternatively, just RCS alone). A sensible next step would be to see if other UK data exist and what it may take to access such data. In addition, relevant non-UK data and expert judgment may also be informative. However, there may be situations where expert judgement cannot be used with confidence e.g. because it is a new scenario. In these and other situations it may be appropriate for HSE to decide to collect new exposure data. The same applies for situations where expert evaluations have been applied to calibrate and/or evaluate these estimates.

Once an initial decision has been reached about priority agents and data sources, some investment will be required for establishing a smooth workflow for data capture and integration. This includes the development of systematic approaches to incentivise data collectors and/or holders in sharing their data with OccECIS. Likely developed approaches should be differential depending on the targeting stakeholder in question. For example, researchers can be motivated in sharing their exposure data by getting access and being able to use the modelled predictions for past and future exposure within their research activities either these are impact or epidemiological assessments. Similarly, for motivating industry and business stakeholders the system could allow them to compare their data after entry with the distributions sourcing from the complete relevant dataset. Additionally, by providing their measurement and contextual data the system could inform through an illustrative process for the most optimal intervention to reduce exposure within their workplace. As far as possible, this should be automated upfront, but it is highly recommended that an ongoing data curator role is created for the maintenance of the extraction regimes and for integration of new data sources into the system. Data sharing activities would greatly benefit if supported by networking activities including presentations in symposia with relevant stakeholders and research conferences, regular newsletters and social media campaigns.

On the statistical modelling side, black-box models are to be avoided for high-impact questions and any calculation of estimates is to be done with the involvement of expert-crafted models, as well as ongoing expert assessments. Calibration and validation of such models is to be encouraged.

Methods for examining trends in exposure data are already established and used by occupational exposure scientists for exposure prevalence or exposure intensity, for example when estimating exposure over time for a JEM in a large epidemiological study. Models for examining trends in risk management measures can be developed using standard regression approaches, or by using expert-crafted Bayesian reasoning approaches, but qualitative approaches may also be possible. The evaluation of interventions at a population level can work on the same theoretical and methodological background that impact assessments for the establishment of OEL are performed upon (e.g. http://www.occupationalcancer.eu/) Criteria for determining which approach should be used in which situation will need to be developed. The sample also applies to the development of uncertainty ranges. Standard methods also exist for estimating bias and study quality. However, the integration of these measures for a scenario is not straightforward and further work is required to clarify the optimum approaches.

Overall, we believe that it is feasible and important to develop an occupational exposure-control intelligence system in order to derive high-level estimates such that HSE and its stakeholders can identify and prioritise hazards and sectors and occupations of concern. We believe that developing a prototype system with RCS, as the working example is the most appropriate initial task.

8. References

1. Cherrie JW. The beginning of the science underpinning occupational hygiene. The Annals of occupational hygiene. 2003;47(3):179-85.
2. Basinas I, Graveling R, Dixon K, Ritchie PA. Developing a data-driven method for assessing and monitoring exposure to dangerous substances in EU workplaces : European Risk Observatory, Summary. Bilbao, Spain: European Agency for Safety and Health at Work; 2019.
3. Kauppinen T, Toikkanen J, Pedersen D, Young R, Ahrens W, Boffetta P, et al. Occupational exposure to carcinogens in the European Union. Occupational and environmental medicine. 2000;57(1):10-8.
4. Peters CE, Ge CB, Hall AL, Davies HW, Demers PA. CAREX Canada: an enhanced model for assessing occupational carcinogen exposure. Occupational and environmental medicine. 2015;72(1):64-71.
5. Scarselli A, Montaruli C, Marinaccio A. The Italian information system on occupational exposure to carcinogens (SIREP): structure, contents and future perspectives. The Annals of occupational hygiene. 2007;51(5):471-8.
6. Fransman W, Schinkel J, Meijster T, Van Hemmen J, Tielemans E, Goede H. Development and evaluation of an exposure control efficacy library (ECEL). The Annals of occupational hygiene. 2008;52(7):567-75.
7. CPWR. ECD: Exposure control database: The Center for Construction Research and Training; 2018 [Available from: https://ecd.cpwrconstructionsolutions.org/.
8. BCCSA. Silica Control ToolTM: Bristish Columbia Construction Safety Alliance; 2021 [Available from: https://www.silicacontroltool.com/.
9. Stefan Hahn S, Blümlein K, Feddersen B, Simetska N, Gillies A, Woolley A, editors. P42 Efficiency of exposure control measures – developing a User Database and Communication Tools. ISES-EUROPE: 1 st European Exposure Science Strategy Workshop; 2018; Dortmund, Germany: ISES-EUROPE.
10. Cherrie JW, Hutchings S, Gorman Ng M, Mistry R, Corden C, Lamb J, et al. Prioritising action on occupational carcinogens in Europe: a socioeconomic and health impact assessment. British journal of cancer. 2017;117(2):274-81.
11. Kauppinen T, Toikkanen J, Pukkala E. From cross-tabulations to multipurpose exposure information systems: A new job-exposure matrix. American journal of industrial medicine. 1998;33(4):409-17.
12. Kauppinen T, Heikkila P, Plato N, Woldbaek T, Lenvik K, Hansen J, et al. Construction of job-exposure matrices for the Nordic Occupational Cancer Study (NOCCA). Acta Oncol. 2009;48(5):791-800.
13. van Tongeren M, Kincl L, Richardson L, Benke G, Figuerola J, Kauppinen T, et al. Assessing occupational exposure to chemicals in an international epidemiological study of brain tumours. The Annals of occupational hygiene. 2013;57(5):610-26.
14. Burns DK, Beaumont PL. The HSE National Exposure Database--(NEDB). The Annals of occupational hygiene. 1989;33(1):1-14.
15. Peters S, Vermeulen R, Olsson A, Van Gelder R, Kendzia B, Vincent R, et al. Development of an exposure measurement database on five lung carcinogens (ExpoSYN) for quantitative retrospective occupational exposure assessment. The Annals of occupational hygiene. 2012;56(1):70-9.
16. Basinas I, Liukkonen T, Sigsgaard T, Andersen NT, Vestergaard JM, Galea K, et al. P096 Statistical modelling and development of a quantitative job exposure matrix

for wood dust in the wood manufacturing industry. Occupational and environmental medicine. 2016;73(Suppl 1):A152-A3.

17. Basinas I, Wouters IM, Sigsgaard T, Heederik D, Spaan S, Smit LA, et al. O46-4 Development of a quantitative job exposure matrix for endotoxin exposure in agriculture. Occupational and environmental medicine. 2016;73(Suppl 1):A88-A.
18. Kauppinen T, Vincent R, Liukkonen T, Grzebyk M, Kauppinen A, Welling I, et al. Occupational exposure to inhalable wood dust in the member states of the European Union. The Annals of occupational hygiene. 2006;50(6):549-61.
19. Burstyn I, Kromhout H, Cruise PJ, Brennan P. Designing an International Industrial Hygiene Database of Exposures Among Workers in the Asphalt Industry. The Annals of occupational hygiene. 2000.
20. Fransman W, Van Tongeren M, Cherrie JW, Tischer M, Schneider T, Schinkel J, et al. Advanced Reach Tool (ART): development of the mechanistic model. The Annals of occupational hygiene. 2011;55(9):957-79.
21. Lavoue J, Joseph L, Knott P, Davies H, Labreche F, Clerc F, et al. Expostats: A Bayesian Toolkit to Aid the Interpretation of Occupational Exposure Measurements. Annals of work exposures and health. 2019;63(3):267-79.
22. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: A Probabilistic Programming Language. Journal of Statistical Software. 2017;76(1).
23. Bingham E, Chen JP, Jankowiak M, Obermeyer F, Pradhan N, Karaletsos T, et al. Pyro: Deep Universal Probabilistic Programming2018 October 01, 2018:[arXiv:1810.09538 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2018arXiv181009538B.
24. Curry E, Freitas A, O'Riáin S. The Role of Community-Driven Data Curation for Enterprises. In: Wood D, editor. Linking Enterprise Data. Boston, MA: Springer US; 2010. p. 25-47.
25. Freitas A, Curry E. Big Data Curation. In: Cavanillas JM, Curry E, Wahlster W, editors. New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe. Cham: Springer International Publishing; 2016. p. 87-118.
26. Cherrie JW, Gorman MN, Searl A, Shafrif A, van Tongeren MJ, Misty R, et al. Health, socio-economic and environmental aspects of possible amendments to the EU Directive on the protection of workers from the risks related to exposure to carcinogens and mutagens at work: Respirable Crystalline Silica. IOM Research Project; P937/8. Edinburgh, UK: Institute of Occupational Medicine (IOM); 2011.
27. Rushton L, Hutchings SJ, Fortunato L, Young C, Evans GS, Brown T, et al. Occupational cancer burden in Great Britain. British journal of cancer. 2012;107 Suppl 1:S3-7.
28. HSE. Work-related Chronic Obstructive Pulmonary Disease (COPD) statistics in Great Britain, 2020. Avvailable at: www.hse.gov.uk/statistics/causdis/. Assessed 25 May 2021. . Buxton, UK: Health and Safety Executive; 2020.
29. ONS U. Labour Force Survey: Office for National Statistics, UK; 2020 [Available from:https://www.ons.gov.uk/surveys/informationforhouseholdsandindividuals/householdandindividualsurveys/labourforcesurvey
30. ONS U. Annual population survey (APS) QMI: Office for National Statistics; 2012 [Available from: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/methodologies/annualpopulationsurveyapsqmi.

31. ONS U. 2011 Census data: Office for National Statistics, UK; 2011 [Available from: https://www.ons.gov.uk/census/2011census/2011censusdata.
32. Eurostat. Structural Business Statistics — Overview: European Commission; 2020 [Available from: https://ec.europa.eu/eurostat/web/structural-business-statistics.
33. ECHA. Registered substances: ECHA; 2021 [Available from: https://echa.europa.eu/information-on-chemicals/registered-substances.
34. ECHA. C&L Inventory: ECHA; 2021 [Available from: https://echa.europa.eu/information-on-chemicals/cl-inventory-database.
35. HSE. EH40/2005 Workplace exposure limits. Norwich, UK: TSO (The Stationery Office); 2020. 60 p.
36. IFA D. Foreign and EU limit values: Institut für Auslandsbeziehungen, Germany; 2021 [Available from: https://www.dguv.de/ifa/fachinfos/occupational-exposure-limit-values/foreign-and-eu-limit-values/index.jsp.
37. OSHA. Permissible Exposure Limits – Annotated Tables. Available at: https://www.osha.gov/annotated-pels Washington, USA: Occupational Safety & Health Administration, United States Department of Labor. Washington; 2021 [
38. IARC. List of Classifications: International Agency for Research on Cancer; 2021 [Available from: https://monographs.iarc.who.int/list-of-classifications.
39. Koppisch D, Schinkel J, Gabriel S, Fransman W, Tielemans E. Use of the MEGA exposure database for the validation of the Stoffenmanager model. The Annals of occupational hygiene. 2012;56(4):426-39.
40. Vincent R, Jeandel B. COLCHIC-occupational exposure to chemical agents database: current content and development perspectives. Applied occupational and environmental hygiene. 2001;16(2):115-21.
41. Vinzents P, Carton B, Fjeldstad P, Rajan B, Stamm R. Comparison of Exposure Measurements Stored in European Databases on Occupational Air Pollutants and Definition of Core Information. Applied occupational and environmental hygiene. 2011;10(4):351-4.
42. Flanagan ME, Seixas N, Becker P, Takacs B, Camp J. Silica exposure on construction sites: results of an exposure monitoring data compilation project. Journal of occupational and environmental hygiene. 2006;3(3):144-52.
43. Zilaout H, Vlaanderen J, Houba R, Kromhout H. 15 years of monitoring occupational exposure to respirable dust and quartz within the European industrial minerals sector. International journal of hygiene and environmental health. 2017;220(5):810-9.
44. Beaudry C, Lavoue J, Sauve JF, Begin D, Senhaji Rhazi M, Perrault G, et al. Occupational exposure to silica in construction workers: a literature-based exposure database. Journal of occupational and environmental hygiene. 2013;10(2):71-7.
45. Peters S, Vermeulen R, Portengen L, Olsson A, Kendzia B, Vincent R, et al. SYN-JEM: A Quantitative Job-Exposure Matrix for Five Lung Carcinogens. The Annals of occupational hygiene. 2016;60(7):795-811.
46. Peters S, Vermeulen R, Cassidy A, Mannetje A, van Tongeren M, Boffetta P, et al. Comparison of exposure assessment methods for occupational carcinogens in a multi-centre lung cancer case-control study. Occupational and environmental medicine. 2011;68(2):148-53.
47. Fevotte J, Dananche B, Delabre L, Ducamp S, Garras L, Houot M, et al. Matgene: a program to develop job-exposure matrices in the general population in France. The Annals of occupational hygiene. 2011;55(8):865-78.

48. Sauve JF, Beaudry C, Begin D, Dion C, Gerin M, Lavoue J. Silica exposure during construction activities: statistical modeling of task-based measurements from the literature. The Annals of occupational hygiene. 2013;57(4):432-43.
49. Peters S, Vermeulen R, Portengen L, Olsson A, Kendzia B, Vincent R, et al. Modelling of occupational respirable crystalline silica exposure for quantitative exposure assessment in community-based case-control studies. Journal of environmental monitoring : JEM. 2011;13(11):3262-8.
50. Tielemans E, Marquart H, De Cock J, Groenewold M, Van Hemmen J. A proposal for evaluation of exposure data. The Annals of occupational hygiene. 2002;46(3):287-97.
51. Zilaout H, Houba R, Kromhout H. Temporal trends in respirable dust and respirable quartz concentrations within the European industrial minerals sector over a 15-year period (2002-2016). Occupational and environmental medicine. 2020;77(4):268-75.
52. Creely KS, Cowie H, Van Tongeren M, Kromhout H, Tickner J, Cherrie JW. Trends in inhalation exposure--a review of the data in the published scientific literature. The Annals of occupational hygiene. 2007;51(8):665-78.
53. Fritschi L, Sadkowsky T, Glass DC. OccIDEAS: web-based assessment of occupational agent exposure. International journal of epidemiology. 2020;49(2):376-9.

Find Out More

Contact us at ashton@manchester.ac.uk

Follow us on Social Media

@AshtonInstitutue

www.linkedin.com/company/thomas-ashton-institute

/ThomasAshtonInstitute

TAI Thomas Ashton Institute