Towards guidelines for time-trend reviews examining temporal variability in human biomonitoring data of pollutants

Author(s):
Sharma, Brij M.; Kalina, Jiří; Whaley, Paul; Scheringer, Martin

Publication date:
2021-06

Permanent link:
https://doi.org/10.3929/ethz-b-000472846

Originally published in:
Environment International 151, https://doi.org/10.1016/j.envint.2021.106437

Environment International 151 (2021) 106437

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Towards guidelines for time-trend reviews examining temporal variability in human biomonitoring data of pollutants

Brij Mohan Sharma [a,*], Jiří Kalina [a], Paul Whaley [b], Martin Scheringer [a,c]

[a] RECETOX, Masaryk University, 62500 Brno, Czech Republic
[b] Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, United Kingdom
[c] Institute of Biogeochemistry and Pollutant Dynamics, ETH Zürich, 8092 Zürich, Switzerland

ARTICLE INFO

Handling Editor: Frederic Coulon

Keywords:
Time trend reviews
Systematic reviews
Human biomonitoring
Pollutants
Guidelines
Research methods

ABSTRACT

In the last few decades, a plethora of studies have focused on human biomonitoring (HBM) of chemical pollutants. Reviewing the copious HBM data reported in these studies is essential for evaluating the effectiveness of pollution management efforts, for example by evaluating time-trends. Nevertheless, guidance to systematically evaluate time trends in published HBM data has never been developed. In this study, we therefore present a proposal for guidelines to conduct "time-trend reviews" (TTRs) that examine time trends in published large HBM datasets of chemical pollutant concentrations. We also demonstrate the applicability of these guidelines through a case study that assesses time-trends in global and regional HBM data on mercury. The recommended TTR guidelines in this study are divided into seven steps: formulating the objective of the TTR, setting up of eligibility criteria, defining search strategy and screening of literature, screening results of search, extracting data, analysing data, and assessing certainty, including the potential for bias in the evidence base. The TTR guidelines proposed in this study are straightforward and less complex than those for conducting systematic reviews assessing datasets on potential human health effects of exposure to pollutants or medical interventions. These proposed guidelines are intended to enable the credible, transparent, and reproducible conduct of TTRs.

1. Introduction

The recent decades have witnessed an exponential growth in scientific publications (Larsen and von Ins, 2010). Alone on the topic of environmental science (miscellaneous), the Scientific Journal Rankings (SJR) displays records of more than 300 journals, book series, and conferences and proceedings, which published approximately 43,000 documents in 2018 (https://www.scimagojr.com/). Moreover, publications in related research areas such as chemical health and safety, environmental chemistry, food science, etc. further add to the existing number of documents in the environmental sciences domain. From such an enormous amount of scientific literature, deriving meaningful conclusions to track the environmental and human body burdens and associated health risks of chemical pollutants is an increasingly challenging task (Haddaway et al., 2018; Whaley et al., 2016).

One approach to deriving reliable conclusions from large bodies of published primary research studies is through systematic review (SR). The concept of SR was originally established in the field of medicine and clinical research to ensure transparency in summarizing evidence (such as dose–response experiments) in a way that minimizes bias and random errors and facilitates assessment of reliability, consensus, and reasons for heterogeneity across relevant studies (Haddaway et al., 2016; Liberati et al., 2009; Schünemann and Moja, 2015). Recently, SR methods have been adopted in a range of disciplines beyond the medical sciences. For example, several types of SR methods (e.g. traditional SRs, systematic evidence mapping, etc.) have been found useful in systematically conducting reviews in the fields of environmental health, epidemiology, and toxicology (Blettner et al., 1999; Hoffmann et al., 2017; Whaley et al., 2016; Wolffe et al., 2019).

One important area where application of SR methods may also be considered is in the assessments of time trends of chemical pollutants, in particular in human biomonitoring (HBM) data (Basu et al., 2018; Frank et al., 2019; Grant et al., 2013; Sharma et al., 2019). Reviews providing time trend assessments of chemicals are now more in demand than ever considering that either the periodic evaluations of several international treaties ("multi-lateral environmental agreements") on chemical pollution management are due or because of the necessity to strengthen regional and national regulations managing chemical pollution and

* Corresponding author.
E-mail address: brij.sharma@recetox.muni.cz (B.M. Sharma).

https://doi.org/10.1016/j.envint.2021.106437
Received 2 November 2020; Received in revised form 29 January 2021; Accepted 30 January 2021
Available online 21 February 2021

safeguarding human health. The need for harmonized HBM data and their systematic assessment is described in the European HBM4EU initiative, which advocates robust approaches to evaluating existing and new HBM data as a solid scientific basis for chemicals policy and risk communication (Apel et al., 2020). Here we denote this type of review that examines time trends in large HBM datasets of chemical pollutant concentrations as "time-trend reviews" (TTRs).

Compared to reviews of studies into human health effects of exposures or interventions, the task of a TTR is different. Typically, reviews that assess time trends in HBM data are based on a large body of literature consisting of up to hundreds of research articles and documents reporting human body burdens of pollutants specific to different regions, populations, analytical methods, etc. (Basu et al., 2018; Frank et al., 2019; Sharma et al., 2019). Importantly, despite this large body of data to be processed, the task of a TTR is generally less demanding than an SR of human health studies. Conducting time trend analyses of HBM data – basically a statistical analysis of a data series – is substantially less complex in terms of methods and data interpretation than in multidimensional and multifaceted datasets on the potential human health effects of environmental challenges or medical interventions. The question of whether a health effect is visible in a large body of data has more serious implications and, technically, requires a more multi-faceted selection and evaluation of data than the question about the "true" value of the time trend in biomonitoring data. For these reasons of data volume and the type of question being asked, we believe that established SR guidelines and approaches cannot be directly applied to TTRs and we propose that, while following the basic principles of SR, specific guidelines for TTRs would be of value to articulate.

In this article, we therefore present a proposal for guidelines for TTRs. In addition, we present a case study of our recently published study on long-term time trends in mercury levels in human tissues (Sharma et al., 2019) in order to illustrate and justify the proposed TTR guidelines. We also use the case study to illustrate how sensitive or insensitive time trends in HBM data may be to an incomplete or biased selection of the HBM data included.

2. Approach to conducting time-trend reviews of HBM data

2.1. Formulating the objectives of the study

Like any other research study, prior to initiating a TTR of a large set of HBM data, it is important to define the research objectives. Once the objectives have been decided, it will help structure and plan further steps of the TTR related to data collection, analysis, and interpretation (see Fig. 1). Essentially, it is advised to define the objective in such a way that it is feasible (in terms of time and resources), original, tractable, of high priority, and contributes to decision making (Vandenberg et al., 2016).

The research objective of a TTR may be to support the implementation of international conventions on pollution management, to improve and/or reorganize local and regional environmental and health policies and priorities, to protect a vulnerable population group (e.g. children, pregnant and breastfeeding mothers, ethnic groups, etc.), to understand the behaviour of pollutants in a human population over a specific time-period, or to assess the quality of existing reviews.

Similar to the PECO (Population, Exposure, Comparator, Outcome) mnemonic used for characterising the objectives of systematic reviews of exposures (Morgan et al., 2018), we recommend formulating the objectives of a TTR according to the TEMPR mnemonic:

- **T**ime Period (over which the trend is occurring)
- **E**xposure (pollutant of concern)
- **M**atrix (the biological matrix or matrices in which the pollutant is measured)
- **P**opulation (in whom the trend and exposure is being investigated)
- **R**egion (the geographical region in which the trend is being investigated)

The decision to select the time period for which the trend analysis is to be performed may vary depending on whether the priority is to compare trends in HBM prior and posterior to a local or regional policy reform (to evaluate the effectiveness of a policy implementation), to assess the changes in lifestyle over a certain period (in terms of usages, environmental release/discharge of a pollutant), or to assess the changes in economic and environmental priorities at the local or global scale that are influencing human exposure to pollutants. The selection of a target pollutant(s) (exposure) for the objective formulation should be based on its current and/or historical usages, environmental and health concerns, and the need of related policy interventions or evaluations. The selection of the target matrix in a TTR is dependent on its suitability for providing sufficiently accurate and extensive measurements of the target pollutant's human body burden. The selection of the matrix also depends on

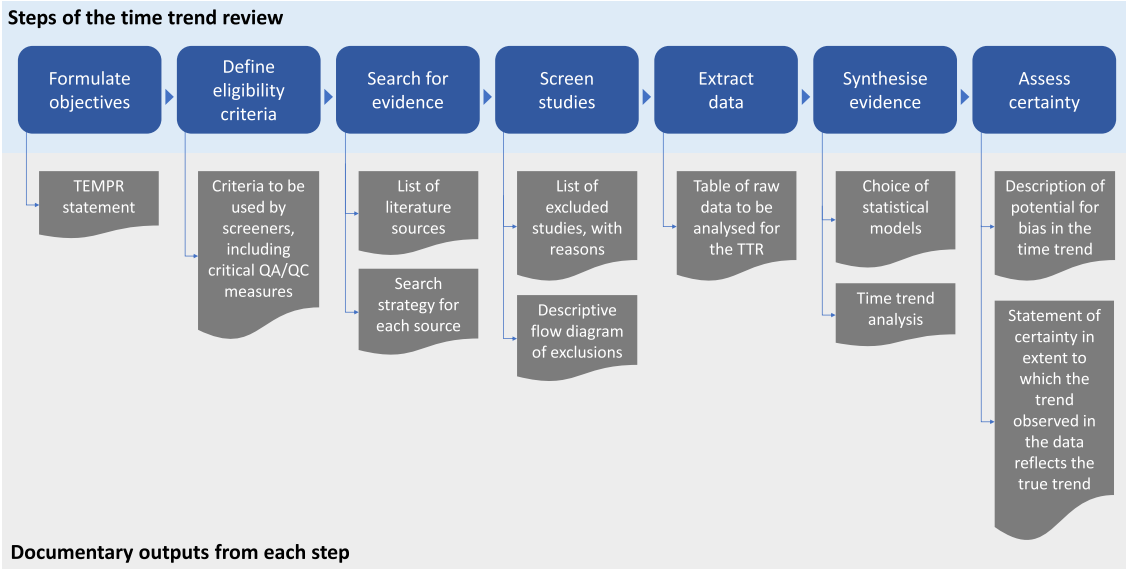

Fig. 1. Summary of the steps and documentary outputs of a Time-Trend Review (TTR).

the selection of the target population type. The target population could be the general population or a specific population in terms of pollution exposure type or vulnerability (in terms of age group, sex, geographic location, ethnicity, etc.). In many cases, the selection of time period, exposure, matrix, and population is strongly influenced by the choice of the geographical region, for example, investigating time trends in pollution exposure in children involved in mining activities in countries from Asia and Africa.

2.2. Eligibility criteria

The eligibility criteria determine the search concepts and strategies as well as the rationale for exclusion of irrelevant studies when screening the search results. The eligibility criteria for a TTR should be based on the elements of the TEMPR statement. These elements can often be directly translated into eligibility criteria, at least for TTRs with straightforward objectives. Eligibility criteria for a TTR can be framed by specifying the time period for which the time trend is to be estimated, exposure (pollutants of concern and exposure pathway), matrix (in which the exposure is to be measured), population characteristics (age, sex, comorbidity, and socio-economic status), and region (geographical location).

In addition to the TEMPR elements, each individual study should be screened for quality assurance/quality control (QA/QC) measures which have a critical impact on the validity of the HBM study findings. Particularly important are issues related to population selection (e.g. in terms of its representativeness), analytical methods (e.g. sample preparation, chemical analysis methods, etc.), and data analysis (e.g. data segregation, statistical methods, etc.). Potential criteria for QA/QC screening include performance in inter-laboratory comparison (ILC) exercises (if the study is part of a large HBM project such as HBM4EU), internal QC measures (in the pre-analytical phase, analytical phase, and quality assurance phase) (Angerer et al., 2007), and the validity of analytical methods e.g. in accordance with the European Medicines Agency (EMA) or the Environmental Protection Agency (USEPA) guidelines. Care should be taken to ensure that application of QA/QC criteria does not bias the dataset or exclude important evidence from the TTR. This may happen, for example, if ILC exercises are made mandatory for a TTR but in fact are conducted only for a subset of studies from the largest coordinated HBM projects.

The purpose of evaluating studies for quality control measures at the screening stage is to prevent them from introducing bias into the overall time trend analysis. This is a major point of difference from conventional SR methods, which would include studies regardless of their validity but only synthesise those studies determined to be of sufficiently high validity (Whaley et al., 2020a). The reason for the difference in approach is that risk-of-bias assessment has relatively limited utility in TTRs, as bias in point estimates in individual HBM studies has relatively little impact when the overall time trend is the target of analysis. Nonetheless, studies with critical issues should be excluded. (TTRs which seek to conduct detailed risk-of-bias assessment should not exclude studies for QA/QC issues. Risk-of-bias assessment in TTRs is beyond the scope of the present guidance.) The criteria for excluding studies according to QA/QC should be defined by the TTR authors depending on the objectives of the TTR and consistently applied to all studies. The studies which have been excluded for QA/QC issues should be recorded.

2.3. Search strategy

The results of a TTR could be biased if the evidence included in the TTR is not representative of the evidence base as a whole. It is challenging to not be inadvertently selective in locating evidence for a TTR. In spite of authors' best efforts, publication bias means some evidence remains inaccessible to researchers, the way research databases are indexed and updated means some relevant studies end up being missed, and uneven accessibility of grey literature means evidence relevant to a TTR can be overlooked (Gusenbauer and Haddaway, 2020a).

A major difference between typical SRs and TTRs is the primary importance in TTRs of grey literature retrieval. While HBM studies reporting human body burdens of pollutants are often published in peer-reviewed research articles, many HBM surveys are performed by national or international public health agencies and published as reports. These reports are considered as non-peer-reviewed (grey) literature, but their quality is often high in terms of study design, analytical methods, and data analysis and reporting. For example, HBM data from IPCHEM, GENASIS, NHANES, KNHANES etc. are part of the grey literature but these comply with quality control and assurance criteria. In addition to the reports from national and international agencies, grey literature from other sources such as dissertations, conference proceedings, regulatory databases, case reports etc. can be considered for inclusion if the quality of their data is confirmed. To assess the quality of the data from grey literature, several characteristics of the data can be checked; these include a well-defined and representative methodology of sampling, analytical procedures including QA/QC, and methods adopted for statistical treatment of data and their consistency in time.

Sometimes, the HBM data generated by national/international public health agencies are simultaneously processed by different research groups (analysing data for different objectives) and side by side published in form of reports and research articles. In such a situation, it is necessary to avoid the duplication of HBM data by selecting the primary study reporting the complete original dataset. This will also help avoid the bias that may occur due to selective outcome reporting.

As our case study shows (see Section 3), the methods used for calculating time trends imply that selective inclusion of evidence may be less of an issue in a TTR. Nonetheless, efforts to include all relevant literature should err on the comprehensive side and documentation of search methods should be complete and transparent, to both minimise risk of bias and ensure that the results of the TTR are reproducible and can be appraised for risk of bias in inclusion of evidence. Because of the complexity of the evidence base for TTRs, in particular the combination of uneven indexing across databases and the need to retrieve grey literature, we recommend that search strategies be planned with an information specialist (Rethlefsen et al., 2015a).

2.3.1. Databases and search engines

There are three principal ways of locating studies relevant for inclusion in a TTR: databases, search engines, and hand-searching. Databases and search engines are suited to "systematic" and "look up" searches, respectively (Gusenbauer and Haddaway, 2020b) while hand searching, the manual screening of the bibliographies of studies eligible for inclusion in the TTR, allows a search strategy to be rounded out with documents that might otherwise have been inaccessible via the first two methods.

It is only databases that allow a reproducible search to be conducted, and they should therefore be the primary source of studies for a TTR. PubMed is an example of one of the major life sciences databases (Frandsen et al., 2019) and Scopus is an example of one of the largest databases of scientific research (Baas et al., 2020). Platforms such as the Web of Science (Falagas et al., 2008) comprise multiple databases access to which is contingent on institutional subscriptions. Although specific subscriptions available at the time of search can be made transparent, differences between institutional subscriptions mean platforms such as Web of Science present challenges for the reproducibility of search strategies; however, these platforms may provide access to studies which are otherwise difficult to retrieve. Search engines such as Google Scholar allow retrieval of documents (Tober, 2011), especially grey literature, which are known to the authors of a TTR; however, these are not databases and do not facilitate a comprehensive or systematic search. For regional biomonitoring, databases such as NHANES (Patel et al., 2016), etc. should be considered for relevant literature searches. Judicious use of databases, search engines, and hand-searchers is a necessary part of a comprehensive retrieval of evidence relevant to the objectives of a TTR.

For a comprehensive review of the suitability of a wide range of academic search systems for supporting systematic reviews, we direct the reader to a study by Gusenbauer and Haddaway (Gusenbauer and Haddaway, 2020b).

2.3.2. Search terms

A major challenge in retrieving evidence for a TTR is that language is variable, with researchers using different terms to describe the same concepts. For example, across a body of literature a chemical may be referred to by several different names, numeric identifiers, and acronyms (Whaley et al., 2020b). Research databases and platforms such as PubMed and Web of Science address this by indexing studies with controlled vocabularies; however, these vocabularies are always finite in scope, indexing is a time-consuming process, and the choices of controlled vocabulary and topics prioritised for indexing all vary across different platforms. Understanding how to search effectively in each platform is therefore a specialist skill which authors of TTRs should not underestimate; engagement with librarians and information specialists is therefore highly recommended (Rethlefsen et al., 2015b). Transparent documentation of search strategies is essential, with the search string used for each database and the date when the search was conducted being vital information which should be provided in the TTR manuscript.

2.3.3. Grey literature

While there is plentiful published guidance in the literature and provided by libraries (Harari et al., 2017; Leenaars et al., 2012; University of Michigan Library, 2020), this kind of systematic search guidance is often designed for the purpose of collecting peer-reviewed publications. In the case of TTRs, where equal importance might be given to grey literature (e.g. institutional reports) for the purpose of retrieving HBM data, the traditional approach to literature search should accommodate modifications. These modifications can be in form of additional search steps added to the existing traditional systematic search approach (Adams et al., 2017; Godin et al., 2015). These additional search steps are inclusion of grey literature databases (e.g. searching dissertations on Embase or Web of Science), customization of world-wide web search engines, and browsing websites of relevant national or regional HBM programmes (such as NHANES, HBM4EU, etc.), organizations (e.g. the World Health Organization (WHO), the German Human Biomonitoring Commission, etc.), and large cohorts (e.g. CHEF research in the Faroe Islands, the CELSPAC study in the Czech Republic, etc.) (Apel et al., 2020; Larsen et al., 2013; Lewis et al., 2017; Patel et al., 2016; Schulz et al., 2007; WHO, 2015). In addition, further grey literature could be retrieved by contacting experts (individuals that are well-versed in the topic of time trend analysis of pollution exposure and likely to be aware of relevant documents) to identify other possible sources for inclusion in the review.

2.4. Screening results of search

Once the comprehensive literature search is completed, the next step in a TTR is to screen the generated citations (and abstracts) against the eligibility criteria defined in the TEMPR statement. This is to identify from all the search results which studies are actually relevant to the TTR objectives. This should be conducted in two steps. In the first, titles and abstracts should be reviewed so that studies that are obviously not relevant to the review objectives can be excluded. Studies for which the relevance is difficult to judge then need to be read in full text and their eligibility is further assessed. Ideally, this should be done in duplicate (Wang et al., 2020).

Review authors should document the screening process in their TTR. The documentation should include how studies were screened (e.g. in duplicate by the reviewers), any critical QA/QC issue which resulted in exclusion (one reason is sufficient), what kind of studies (and how many) were excluded/included in each step of screening, and what was the major reason for exclusion of studies within the two-steps of screening. A descriptive flow diagram should be used to summarize the screening process in a TTR (see Fig. S1 in the Supplementary Material (SM)) (Page et al., 2020).

2.5. Data extraction

Data extraction is the process of collecting relevant information from the full-text version of the finally selected studies for the subsequent data analysis step (Hoffmann et al., 2017). TTR authors should plan what data are required and relevant for the analysis of time trends in HBM. The data to be extracted should be adequate to summarize the collected studies, correspond to the objectives of the TTR, and enable complete data analysis (ideally, even including assessment of risk of bias). For a TTR, the most relevant data to be extracted from each selected study can be grouped under three categories: mandatory, highly recommended, and recommended (presented in Table 1).

The list of specific features to be extracted from each study should be conceived at the beginning of the data extraction process. The extracted data should be entered in a suitable software. There are specialist software packages which, although originally developed for systematic review, can be adapted for data extraction and other steps of the planning and conduct of the TTR process. Kohl et al. present a comprehensive review of these existing online tools (Kohl et al., 2018). Spreadsheet software is also often used for data extraction; however, extracting study data directly into a spreadsheet can increase risk of issues such as transposition errors, so use of data extraction forms is recommended.

During the process of data extraction one major challenge is the inhomogeneity of the extracted data. For example, within the information about population characteristics, different studies may follow different cut-offs to define age groups (infants/children/adolescents/adults/elderly). In certain cases when the primary data in the studies selected for the TTR are available, this inhomogeneity can be controlled by adopting the standard cut-offs recommended by international/regional health organizations (e.g. WHO). If the original study does not explicitly provide primary data, adopting such international cut-offs may not curb the inhomogeneity. Another example of inhomogeneity is different ways in which the descriptive statistics are presented, i.e., different methods to present the central value (mean or median) and variance of the distribution of pollutants concentrations in the reported data, etc. This mix-up of descriptive statistics can be avoided if the primary data from the selected study are made available and could then be used to estimate/

Table 1
List of data items to be considered for collection in TTR.

Data category	Description of data (information)
Mandatory	• Timing and duration of study • Exposure type (or source of exposure: occupational, dietary, background, etc.) • Matrix (pollutant concentration and behaviour (uptake, elimination half-life) in the targeted biological matrix) • Population type (whether it is a general population or a subgroup (based on any specific character including sex, age, ethnicity, etc.) • Region (country/state/city) • Study metadata (authors, year of publication, etc.)
Highly recommended	• Study and sampling design (eligibility criteria, sampling procedures, timing of measurements, etc.) • General characteristics of the sampled population (physiological, socio-economic, age, sex, comorbidity, etc.) • Whether the study is part of a large human-biomonitoring survey • Methods of data aggregation • Statistical descriptors and methods • Analytical QA/QC
Recommended	• Information about any interventions (direct or indirect) (if any) • Control groups (if any) • Reference to other relevant (or parallel) studies

model data values of the reviewers' interest. In addition, inhomogeneity might be also introduced in the extracted data by different units used in analytical measurements or by differently defined demography of populations. Depending on the scale of the TTR (global, regional, or local), in some cases, it might not be possible to exclude factors of inhomogeneity in the extracted data; in such cases effects of inhomogeneity on resulting time-trends should be acknowledged and discussed.

2.6. Analysis of HBM data

Prior to performing the data analysis on the extracted HBM data, several typical properties and patterns of HBM data should be checked and considered as opportunities for a better understanding of the extracted data.

HBM data can follow any statistical distribution, nevertheless most of the HBM datasets (specifically of an individual population group) are log-normally distributed (or close to it) (Albertini et al., 2006). This means that after log-transforming the data, the dataset may be further assessed by statistical methods suitable for normally distributed data. In other cases, when the HBM data do not follow either a normal or log-normal distribution, more advanced (non-parametric) statistical methods are needed to analyse the data.

2.6.1. Methods for estimating time-trends in HBM data

In TTRs, the set of available statistical methods is relatively specific. Preferably the established parametric method of ordinary least-squares linear regression should be used for analysing time trends in HBM data.

The biggest advantage of this method is that it is known to be the best unbiased linear estimator of the series if three conditions are fulfilled. These conditions stated by the Gauss–Markov theorem ("Gauss–Markov Theorem," 2008) (zero mean, finite variance, and no mutual correlation) are met when the residuals of the trend (e.g. non-explained variance) follow a Gaussian (normal) statistical distribution with zero mean. This should always be tested for by using one of the available tests of normality (Anderson-Darling test, Kolmogorov-Smirnov test, and Shapiro-Wilks test (Anderson and Darling, 1952; Kolmogorov, 1933; Shapiro and Wilk, 1965; Smirnov, 1948)). If the residuals are log-normally distributed, a log-transformation of the primary data helps to meet the conditions and the back transformation provides exponential trends rather than (unrealistic) linear ones. If the data do not meet the conditions of the Gauss–Markov theorem, one of the non-parametric methods for trend estimation should be used. The most often used and well described ones are the Theil-Sen single median and Siegel repeated medians trend methods (Sen, 1968; Theil, 1992; Matoušek et al., 1998).

To assess the influence of the sources of bias (described in Section 2.7) on the characteristics of the final trend, a bootstrapping method (e.g. Monte Carlo) should be used. Every source of uncertainty (inappropriate descriptive statistics, approximate or semi-quantitative values, inaccurate normalization, small size of dataset, etc.) can be simulated by a defined random variable. A certain number of realizations of such random variables can then be generated (usually at least tens of thousands of realizations are used depending on the variance in the HBM data collected for trend analysis). The optimal number of realizations additionally depends on the required accuracy of the computation and the available time and resources for computation (Huth, 1999; Zhang et al., 2004). The trend assessment is then performed for each of these realizations and the set of results makes it possible to estimate the range of variability of the final result. Because the analytical assessment of more complex datasets can be complicated, bootstrapping methods are often easier or the only possible way to assess the robustness of the results of the time trend analysis.

2.7. Certainty and bias

2.7.1. Certainty assessment

Because a TTR is developed by synthesising data from existing evidence, and this evidence can be of varying quality, it is important to assess the impact of the quality of the evidence on certainty in the findings of the TTR, i.e., evaluate the extent to which the trend derived from the existing evidence is likely to reflect the "true" trend.

In systematic reviews, one of the most widely used approaches to assessing certainty in the evidence is the GRADE Certainty of Evidence Framework. It assesses certainty in the evidence according to five factors which reduce certainty and three factors which increase it (Guyatt et al., 2008). The framework is being adapted for use in systematic reviews of exposures (Morgan et al., 2019, 2016). We believe that at least some of the general principles of GRADE are applicable to TTR studies as well. We therefore recommend the following issues to be considered when assessing certainty in the results of a TTR; further research in this area should be conducted.

Here, "certainty in the evidence" means certainty that the identified time trend is true. The lower the certainty is, the more likely it is that if additional studies are conducted that address one or more of the issues identified below, the time trend would change.

The factors which may reduce certainty in the evidence collected for TTRs are as follows:

- **Potential for bias**: Limitations in the design and conduct of HBM studies can bias the point estimates of exposure upon which time trend analyses are based. The more serious and widespread the limitations are across the evidence base, the lower the certainty in the time trend. We discuss these in more detail later in Section 2.7.2.
- **Inconsistency of results of studies informing the TTR**: If the studies on which a time trend analysis is based have differing results for reasons which cannot be explained by differences in how they were designed or conducted, then it is less clear where the true trend actually lies.
- **Indirectness of the evidence**: The target population for a TTR may be different from the actual populations which are the subject of the studies included in the review. The more important those differences are, the less certainty there is in the time trend.
- **Imprecision**: Imprecision refers to the width of the confidence intervals around a time trend. The wider the intervals, the less certainty there can be as to where the true trend lies; if the confidence intervals cross the null, then there may be no trend at all.
- **Publication bias**: Publication bias arises from selective publication of studies based on their results. This distorts the evidence which is available for review. The more concerns there are that publication bias is affecting the studies included in a TTR, the lower the certainty in the time trend.

The factors that raise certainty in the results of a systematic review include large magnitude of effect, the presence of a dose–response gradient, and plausible residual confounding tending towards the null hypothesis (Morgan et al., 2019). Of these, only magnitude of effect seems directly applicable to TTRs, whereby a steep trend should overwhelm concerns about residual biases, provided there are not already concerns about bias in the evidence for the TTR.

Operationalising the assessment of certainty in the evidence in the context of conducting TTRs is an issue which should be further explored. Pending such operationalisation, researchers should assess certainty in TTRs by carefully considering the potential impact of the five factors which reduce certainty (potential for bias, inconsistency, indirectness, imprecision, and publication bias), plus magnitude of the trend, on the time trend analysis.

2.7.2. Common potential sources of bias in TTRs

There are several sources that may introduce random uncertainty or

create a bias (systematic shift) in trends in HBM data. Whereas most of the uncertainty sources are common for the analysis of biological data in general, a trend analysis result based on many data points may be less prone to bias (and uncertainty from lack of precision) than other outcomes such as descriptive statistics or difference tests. This robustness may origin from either a smoothing of random error by using multiple points or from a low importance of the systematic bias for some of the trend characteristics (e.g. a relative change in time) as mentioned further below.

The uncertainty/bias in TTR may be present in one or several data points only or may arise from differences between the multiple points. In this section we describe the influence of individual uncertain/biased studies as well as the uncertainty/bias of the overall trend caused by aggregating multiple certain/unbiased points with limited comparability. Here we do not focus on possible uncertainty/bias sources within individual studies as this is generally beyond the scope of a TTR. We nonetheless emphasise that, when assessing potential for bias in a TTR, each of these common issues should be assessed in turn, and any additional sources of bias be carefully considered.

2.7.2.1. Long-term seasonality. Although long-term HBM data are usually not very prone to reveal seasonal fluctuations, it should be tested whether e.g. nutrients or pollutant intake and/or its decomposition is related to seasonal changes and associated factors such as temperature, rainfall, length of sunshine, harvest periods etc.

2.7.2.2. Short-term fluctuations. Short-term fluctuations in the concentration of a pollutant in a biological sample may occur when the pollutant has a short half-life in the organism in which it is being measured, for example fluctuations (within a day) in phthalate concentration in human urine samples. In such cases, pooling multiple samples or adjusting for periodicity should be considered.

Concentrations of pollutants with short elimination half-lives can fluctuate widely in an individual over periods of time as short as days or hours. The level of fluctuation depends on the physicochemical properties of the chemicals and the physiology of the human body (e.g., variation in urinary volume or variation in creatinine excretion rates). Spot sampling of such compounds leads to very imprecise results at the individual level. To compensate for this, TTR authors can restrict their criteria to select studies that have used multiple-spot sampling with statistical averaging or physically pooling of several samples covering the whole sampling period; otherwise, an adjustment for periodicity should be considered (Aylward et al., 2012). In some cases, the HBM data in a target matrix can be adjusted for periodicity by using physiological parameters. For example, for creatinine adjustment the analyte concentration is divided by the creatinine concentration. The reliability of such adjustments can be evaluated by correlating the adjusted data with data measured in the matrix considered (Barr et al., 2005).

2.7.2.3. Left censoring. One typical problem with HBM data may be that a dataset contains a certain number of very low concentrations (below the limit of quantification or detection), which means that the data are semi-quantitative. Usually, the values in HBM data are below the upper quantification limit of the analytical methods, whereas there is no limitation for the lower (i.e., "left") side and the concentration of the target pollutant can be arbitrarily small. The analytical sensitivity of a given method determines what portion of the analysed data will be left censored, i.e., semi-quantitative ("less than").

Omitting the semi-quantitative values would bias the time trend results and should be avoided during the data analysis process. Instead of omitting semi-quantitative data, these data should be substituted by extrapolated values. There are established methods for extrapolating semi-quantitative data such as by replacing the LOQ by LOQ/2 or LOQ/$\sqrt{2}$ or, preferably, by using maximum likelihood estimation methods (Finkelstein and Verma, 2001; Hornung and Reed, 1990). The choice of an extrapolating method depends on the portion of the data below the LOQ, the magnitude of the LOQ, subsequent computations, and established practice in the given research area. Such extrapolation of semi-quantitative data may introduce negligible (in the case of non-parametric methods) or small (in the case of parametric methods and substantially low quantification limits) bias into the trend characteristics.

Small (<10%) portions of left-censored data should not have a significant impact on trends in HBM data (Gilliom et al., 1984) if a proper method of treatment is used (often individual values in left-censored data are replaced by a specific value derived from the statistical distribution of the available data). A greater percentage of left-censored data could potentially introduce a significant impact on trends (in terms of both slope and significance) as well as on other results of the statistical treatment (e.g. descriptive statistics). An assessment of the impact of left-censored data on overall time trends is provided in Section 3 of this paper. Nonparametric techniques such as the Theil-Sen trend estimator can provide unbiased results even when the HBM data contain up to approx. 30% left-censored data (Akritas et al., 1994; Onofri et al., 2019). If the share of left-censored data is higher than 30%, this may lead to rather qualitative than quantitative results of the trend analysis (a concentration decrease is identified but its magnitude is not exactly known). Nonparametric tests such as the Mann-Whitney test of difference (Fisher, 1921; Mann and Whitney, 1947) can identify differences between stratified data subsets up to about 75% of left-censored data (Halperin, 1960).

In cases where the data come from different studies, the value of detection/quantification limits are usually also different. In such cases the substitution of left-censored data should be done with respect to the different LOQs. Moreover, if the left-censored data are abundant, even a properly performed substitution can lead to artifacts and bias on accuracy (e.g. a time series with high LOQ in its initial and low LOQ in its final part can exhibit a false decreasing trend due to the changing level of the substituted values).

2.7.2.4. Selection bias. Selection bias in a TTR may originate from unrepresentative selection of studies and/or from unrepresentativeness present within individual studies of interest. The latter source of bias is generally difficult to handle. Further, selection bias can emerge from the time periods present in the dataset and the stratification of the population covered within individual studies (the population is generally stratified based on socioeconomic conditions, geographic location, age, etc.). The distribution of the stratified groups (i.e., the distribution of the studies) in the compiled HBM dataset should be the same as it is in the general population of interest. Stratification of HBM data into groups should cover all factors possibly affecting the assessed quantity (typically sex, age, living standards, potential exposure etc.).

A difference in a pollutant's concentrations between these groups should be tested by using either a parametric test (Student's *t*-test, ANOVA (Fisher, 1921; Student, 1908)) or some of the nonparametric tests of differences (e.g. Mann-Whitney test, Wilcoxon tests, Kruskal-Wallis test (Kruskal and Wallis, 1952; Mann and Whitney, 1947; Wilcoxon, 1945)). Later, different statistical tests for estimating time trends (as specified in Section 2.6) can be used to identify an overall trend for all groups together.

Nevertheless, both an incorrect time trend and an incorrect difference of time trends between population groups may emerge when the distribution of samples is proportional to the population groups but irregular in time (Simpson's paradox (Simpson, 1951)). To prevent this, it is recommended to calculate partial trends within each of the defined groups and compare differences in the resulting trends. The method of Z-scores (Fisher, 1915) can be used for comparing trend slopes. If there is no significant difference between these partial trends and the overall trend, selection bias is probably negligible.

A special case of selection bias can emerge if the analysis is carried

out on particularly exposed groups (e.g. occupational exposure, locally increased pollution etc.) or if there is an undiscovered factor changing over time (e.g. local migration, income changes, etc.) affecting the data. Usually, such bias is hard to identify and a possible way to avoid it is via careful planning of the review design to ensure that there are no undiscovered exposures (i.e., paying attention to all known circumstances (e.g. occupation, residential area, socioeconomic conditions, etc.) which could represent a potential exposure). Also, the method of partial trends testing over the overall trend can help to identify such potential bias.

2.7.2.5. Analytical bias. Especially changes in instrumentation and/or methods for analysis of a chemical pollutant of interest may cause analytical bias (different devices, reference compounds, estimation of limit of quantification/detection, extraction techniques, measures to eliminate/minimize cross-contamination in laboratory and transport of samples, etc. used in different studies and different time) (Farzanfar et al., 2017). If there is one analytically biased group of data points then this type of bias could be avoided by using appropriate statistical tests (i.e., testing individual trends (based on specific analytical technique, method, etc.) against the overall trend). If the same type of analytical bias is present in all the studies, its influence on the resulting time trend is usually rather low. To identify biased data and minimise potential analytical bias, it is recommended to carry out a detailed investigation of QA/QC criteria and the extent of method validation adopted in the HBM studies collected. For this purpose, different phases of the analytical procedures used in the studies selected to generate HBM data can be evaluated (Angerer et al., 2007). These phases include:

- Pre-analytical phase (time of sample collection, changes in analyte concentration by degradation or evaporation, external contamination during storage and transport, etc.).
- Analytical phase (sample preparation, clean up, calibration, instrument parameters, etc.).
- Quality assurance phase (accuracy of the measurement tools which is usually assessed using control/reference materials for chemicals as well as biological matrices).

Furthermore, analytical bias can be reduced by selecting studies with sufficiently similar analytical methods (or the data sources based on these methods).

A subtype of the analytical bias may be due to the limited comparability of the physicochemical characteristics of the HBM data, e.g. different concentration units (kg vs. L of a liquid matrix; concentration per wet weight or per lipid weight), exact definition of chemical pollutants (racemic mixtures vs. defined isomers) etc. In many cases recalculation methods such as normalization by a given biomarker (e.g. creatinine in urine) are used to make HBM data compatible across samples/matrices/analytical methods but the comparability should also be tested using appropriate statistical tests (i.e., testing individual trends against the overall trend). Such normalization (or adjustment) of analyte concentrations, particularly in urine, is commonly carried out for non-persistent chemicals (i.e., chemicals with short biological half-lives) by dividing the analyte concentration (micrograms analyte per liter urine) by the creatinine concentration (grams creatinine per liter urine). Analyte results are then reported as weight of analyte per gram of creatinine (micrograms analyte per gram creatinine) (Barr et al., 2005; Lermen et al., 2019). TTR authors should explore such possibilities based on solid assumptions, to harmonize HBM data collected from different studies.

2.7.2.6. Statistical bias. A common problem in TTRs may come from different ways of reporting statistical characteristics in selected individual studies. For example, the arithmetic mean and standard deviation are frequently used to report the central value especially in older studies, but often they are not appropriate estimates of the central value (especially in common cases of right-skewed distributions where the mean provides a positively biased estimate of the central value). Depending on the nature of the measured data in primary studies, the geometric mean, different quantiles including the median or other characteristics of the central value should be used to describe the data.

In general, it is not appropriate to combine different descriptive statistics (e.g. central value, variance, etc.) without the knowledge of the statistical distribution of their primary values (i.e., raw data). For symmetric distributions (e.g. the normal distribution) of the primary data, median, arithmetic mean and geometric mean are close and are all suitable estimates of the central value. For log-normal distributions, only the median and geometric mean are appropriate and comparable statistics (Limpert et al., 2001; Shih and Binkowitz, 1987). In more complex cases (e.g. if primary studies reveal different distributions in their data), a stochastic method should be used during trend assessment to avoid statistical bias. Based on known descriptive statistics of the primary data, a new dataset can be generated ("re-sampled") for each individual primary study that fits these characteristics (fitting the distribution to reported statistics such as central values and quantiles using maximal likelihood estimates or other methods). From these datasets generated by re-sampling, any of the above-mentioned descriptive statistics can then be computed, serving as a robust basis for the trend computation itself (Heffernan et al., 2014; Mary-Huard et al., 2007).

3. Case study: long-term time trends of mercury in human tissue

Globally, there exist a few long-term HBM surveys and data on chemical pollutants (WHO, 2015). Mostly, these surveys are specific to regions and countries in Europe and North America. For such comprehensive HBM data, it is usually possible to assess temporal trends of chemical pollutants in human populations and draw reliable conclusions. On the contrary, in regions and countries especially in Africa, Asia, and South America, where chemical pollution issues are frequent, long-term HBM studies of chemical pollutants are scarce due to several reasons related to economy, infrastructure, knowledge resources, etc. (Weiss et al., 2016). Moreover, the available information on HBM of chemical pollutants from such regions is often inhomogeneous in terms of the population's characteristics, temporal distribution, etc. (Barnett-Itzhaki et al., 2018; Basu et al., 2018; Sharma et al., 2019, 2014). Assessment of fragmented and inhomogeneous HBM data is a challenging task. In the previous section we have described steps to collect and analyse large HBM datasets, specifically for the purpose of assessing time trends in a pollutant's concentrations in the human body.

In this section, using a case study on mercury levels in human tissue (Sharma et al., 2019) we illustrate the application of some of the previously described steps in conducting a TTR, and compare time-trends obtained from three scenarios: (i) a scenario where all available data are used and an attempt to follow Systematic Review guidelines – in the absence of TTR guidelines – was made in the acquisition and evaluation of the data (full dataset). (ii) A second scenario where a part of the data was systematically omitted from the time trend analysis. Systematic omission of some data here means complete exclusion of a population sub-group from the full dataset (and, thereby, generating a new reduced dataset). (iii) In the third scenario, a certain fraction of the datapoints (from 1% to 80%) was removed randomly from the full dataset. The last two scenarios simulate situations in which the literature search may not identify and include several relevant studies in the data search and analysis. Furthermore, we offer recommendations for appropriate types (in terms of sample size, population diversity, target pollutant, matrix sampled, etc.) of HBM data.

3.1. Description of the case study

A review (Sharma et al., 2019) assessing time trends of total mercury (THg) in human blood and breast milk samples over a period of 50 years was used to generate the three scenarios. The assessment of time-trends

in this study was based on HBM data obtained from a total of 558 peer-reviewed studies and survey reports. Although this study did not rigidly follow established standard SR guidelines, literature search, selection, and data extraction and analysis were conducted in a systematic and comprehensive fashion. The large HBM dataset included in the study presents several types of inhomogeneity in terms of the populations' characteristics (age, sex, occupation, ethnic background, geographical location, etc.), temporal distribution, analytical methods, etc.

3.2. Method

One of the fundamentals of SR methods is to conduct a systematic search of the literature. The systematic search of the literature involves a detailed and comprehensive search plan and a strategy derived a priori, with the goal of reducing bias by identifying, appraising, and synthesizing all relevant studies on a particular topic (Uman, 2011). In addition to the systematic search, SR methods involve various additional steps (e.g. quality control of the obtained literature, following specific inclusion/exclusion criteria, etc.). In a TTR, however, application of all these steps is often not feasible because of the high number of studies included (several hundreds). We assume that, if the search strategy was not comprehensive enough, the pool of originally collected literature may miss a complete population sub-group (systematic exclusion) (scenario II) or miss (randomly) up to 80% of the collected HBM data (scenario III). The exclusion of some of the HBM data may either represent a non-comprehensive literature search or it may also be a proxy for inconsistent inclusion/exclusion criteria when a systematic search protocol was not strictly followed. Comparing time trends in these datasets, full and reduced (scenarios II and III), allows us to estimate whether the time trends in the HBM data based on a non-comprehensive literature search are different from those in HBM data collected by using a comprehensive search. To determine the difference in trends in these HBM datasets (full and reduced), a p-value of a test of difference (the method of z-scores on Theil-Sen (Sen, 1968; Theil, 1992)) between them was used. In the case of the systematically reduced dataset (scenario II) we obtained one p-value (one reduced dataset), in the case of the randomly reduced dataset (scenario III) we obtained one thousand p-values from which the 5th percentile was chosen.

In a separate analysis, we also addressed the problem of left-censored data. When many (hundreds) of HBM studies from many years or even decades are collected, a particular concern is that there will likely be studies with higher limits of quantification (LOQ), which could influence the time trends obtained. In the full dataset of our case study, the time-trend analysis was based on aggregated data from each study (central values), but data points below the LOQ (left-censored data) are present in the primary data of at least some of the individual studies, which are usually not known. To check the influence of left-censored values, a new dataset was created for each study, based on known characteristics of the data such as total number of participants, median, mean, quantiles etc. These characteristics were available for several individual studies and were used for the re-sampling of artificial primary data for all studies. In these new datasets derived from re-sampling for each individual study, all values below LOQ were substituted by 12.5%, 25%, 50%, and 100% of the LOQ (the LOQ was known for most of the studies). In this analysis, this generation of artificial datasets was repeated 10,000 times to obtain a wide spectrum of estimated trends and p-values of their differences from the trend in the original full dataset. Finally, the 5th percentile of these p-values was taken as an estimation of a significance of the difference between the overall trend and trends computed on individual values including the left-censored ones.

3.3. Results

Scenario I presented in this case study illustrates original time-trends in THg levels in whole blood, cord blood, and breast milk, based on full datasets retrieved from the literature through a comprehensive and systematic search. In all three biological matrices, significant declines in THg levels were observed over the period of about 50 years. More details on results related to scenario I can be found elsewhere (Sharma et al., 2019). In scenario II, we found that in most of the cases a complete exclusion of any of the population sub-groups did not create a significant difference between the time trends computed on the full and reduced datasets (Fig. 2, scenario II: illustration of systematic exclusion of data by excluding all datapoints from a region). Similarly, in scenario III random exclusion of up to 50% of the original HBM data (randomly simulating a non-comprehensive retrieval of data from the literature) did not significantly change the original time trends (Fig. 2, scenario III). In some cases (rich datasets in terms of highly significant trends or large number of data points), even the exclusion of 80% of the original HBM data did not strongly affect the time trends (Fig. 2).

We also found that, under the conditions of our case study where there are mostly high differences between the LOQs and the majority of the values (one order of magnitude), left-censored values had no significant influence on the time trends (see Fig. S2 in the SM). Nevertheless, in some cases, particularly when the HBM dataset is not large enough and excluding data below LOQ is expected to influence the time-trend results in the TTR, it is advised to appropriately treat the data below LOQ. Given that over a period of time HBM data of a chemical pollutant follow a log-normal distribution, left-censored (below LOQ) data can be estimated by using a maximum likelihood method or by substituting LOQ values with $LOQ/\sqrt{2}$ or $LOQ/2$ (Finkelstein and Verma, 2001; Hornung and Reed, 1990).

4. Perspective and way forward

With this work we intend to provide a starting point for a discussion of requirements and guidelines for time-trend reviews (TTRs) based on HBM data of chemical pollutants. Given the increasing need for TTR studies, we think guidelines will be useful in order to make TTR studies as informative, consistent, and reliable as possible.

A first point is that the search for relevant studies needs to be as comprehensive as possible, and the search terms used need to be reported in a way that makes the search transparent and reproducible. In the area of HBM data, a comprehensive search may lead to very large number of studies to be included (hundreds of studies). In such a situation, it is important that simpler (than standard SR) methods for study evaluation and data interpretation make the task tractable and enable authors to inspect and analyse a huge body of literature within reasonable time.

The example of the mercury case study described above represents a data-rich case with many data points clearly above the LOQ. In such a case, a literature search missing a complete population sub-group or up to 20% of the relevant studies is likely to still provide a reliable estimate of the true time trend. Also, the effect of left-censored data is, under these conditions, not strong and a time trend derived from data points reported as average or mean concentrations is likely to be reliable. However, in cases where the number of data points is smaller and the data are closer to the LOQ, the different sources of bias or inhomogeneity in the data may affect the time trends substantially and their influence needs to be evaluated carefully: the smaller the number of studies and data points available and/or the higher the heterogeneity of the data from different periods of time and/or the higher the scatter in the data, the more difficult it will be to determine a reliable time trend. It may still be possible to determine a time trend, but careful evaluation of the available data and of the confounding factors will be needed to determine its significance. Applying stricter data selection criteria might lead to exclusion of a larger fraction of data points but also to higher quality/consistency of the data points included. This may make the time trends either less significant (fewer data points) or more significant (fewer outliers and biased data points). On the other hand, applying more lenient selection criteria may result in a richer dataset and thus the time trends can again be either less significant (more scatter) or more

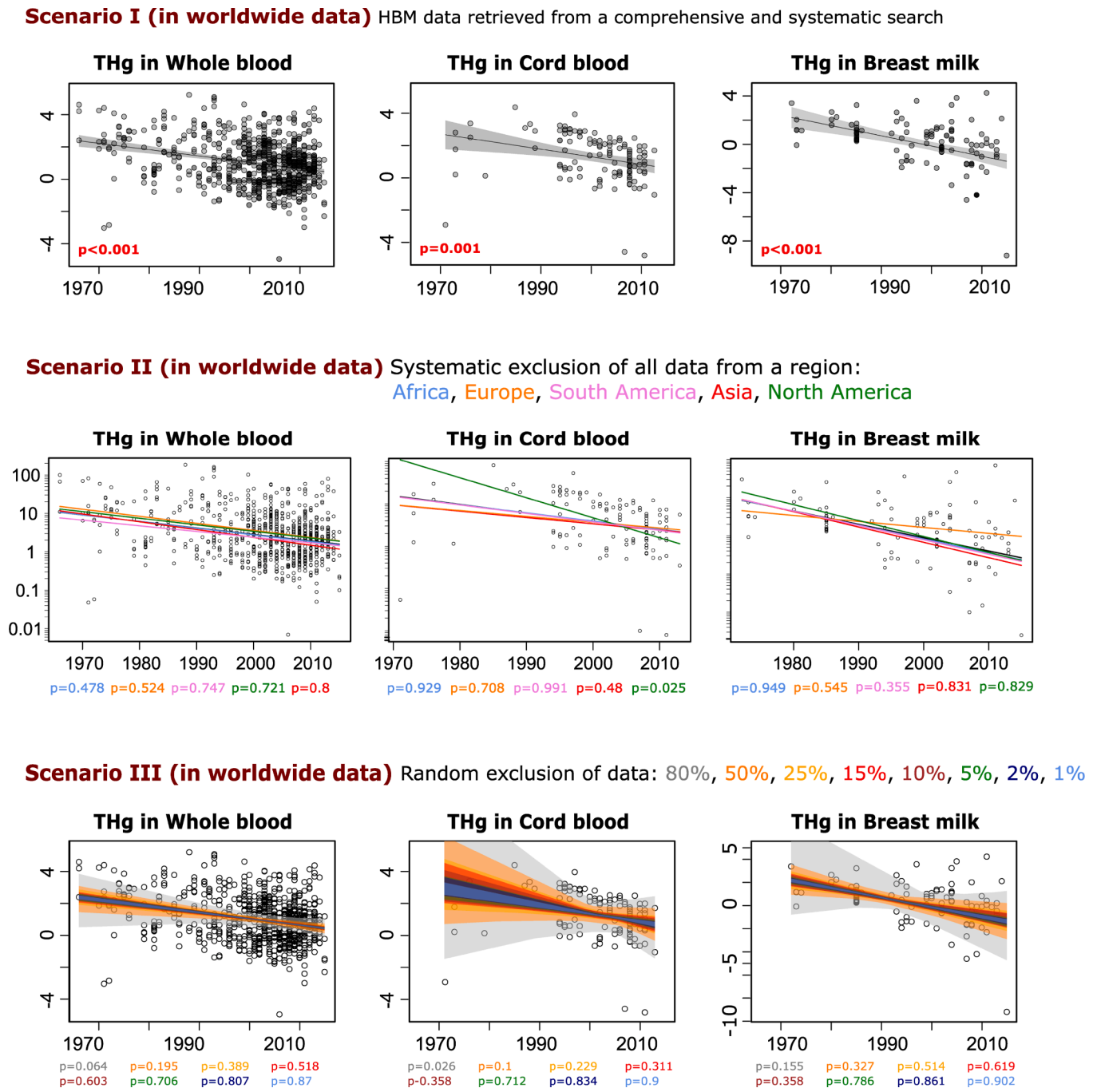

Fig. 2. Illustration of time-trends in the full (scenario I) and reduced HBM datasets (scenarios II and III). p-values provided in scenario I indicate the significance of trends. p-values provided in scenarios II and III indicate if the trends are the same or different compared to the trend from the full dataset. The panels for scenario I show original time trends (in the full dataset retrieved through a comprehensive and systematic search) of THg levels in whole blood, cord blood, and breast milk. The panels for scenario II show new time-trends (generated after predefined exclusion of a stratified population sub-group from the full dataset, in this case population groups from a specific region) in comparison with the time-trends of the full dataset. The panels for scenario III show the time trends computed for the reduced datasets with random exclusion of data points (1%, 2%, 5%, 10%, 15%, 25%, 50% and 80%). For an extensive presentation and analysis of the time trends in the original HBM dataset (full dataset), see Sharma et al. (Sharma et al., 2019).

significant (more data points).

With the guidelines proposed here, we want to initiate a broader discussion of the requirements for TTR studies. For the development of an agreed-upon and widely adopted set of guidelines, more case studies that focus on different types of chemicals and biological matrices and thus investigate different sources of bias and inhomogeneity and how they affect time trends will be highly valuable.

Credit authorship contribution statement

Brij Mohan Sharma: Conceptualization, Writing - original draft. **Jiří Kalina:** Writing - original draft. **Paul Whaley:** Writing - original draft. **Martin Scheringer:** Conceptualization, Wrting and Revising - original draft, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was also supported by the Operational Programme Research, Development and Innovation – the CETOCOEN PLUS project (CZ.02.1.01/0.0/0.0/15_003/0000469). BMS, JK, and MS also thank the Research Infrastructure RECETOX RI (LM2018121) project financed by the Ministry of Education, Youth and Sports, and the Operational Programme Research, Development and Innovation – project CETOCOEN EXCELLENCE (CZ.02.1.01/0.0/0.0/17_043/0009632) for supportive background.

Disclaimer

The views expressed in this article do not necessarily represent the views of authors' affiliations.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2021.106437.

References

Adams, R.J., Smart, P., Huff, A.S., 2017. Shades of grey: guidelines for working with the grey literature in systematic reviews for management and organizational studies. Int. J. Manage. Rev. 19, 432–454. https://doi.org/10.1111/ijmr.12102.

Akritas, M.G., Ruscitti, T.F., Patil, G.P., 1994. Statistical analysis of censored environmental data. Handb. Stat. https://doi.org/10.1016/S0169-7161(05)80009-4.

Albertini, R., Bird, M., Doerrer, N., Needham, L., Robinson, S., Sheldon, L., Zenick, H., 2006. The use of biomonitoring data in exposure and human health risk assessments. Environ. Health Perspect. 114, 1755–1762. https://doi.org/10.1289/ehp.9056.

Anderson, T.W., Darling, D.A., 1952. Asymptotic theory of certain "Goodness of Fit" criteria based on stochastic processes. Ann. Math. Stat. 23, 193–212. https://doi.org/10.1214/aoms/1177729437.

Angerer, J., Ewers, U., Wilhelm, M., 2007. Human biomonitoring: state of the art. Int. J. Hyg. Environ. Health 210, 201–228. https://doi.org/10.1016/j.ijheh.2007.01.024.

Apel, P., Rousselle, C., Lange, R., Sissoko, F., Kolossa-Gehring, M., Ougier, E., 2020. Human biomonitoring initiative (HBM4EU) – strategy to derive human biomonitoring guidance values (HBM-GVs) for health risk assessment. Int. J. Hyg. Environ. Health 230, 1438–4639. https://doi.org/10.1016/j.ijheh.2020.113622.

Aylward, L.L., Kirman, C.R., Adgate, J.L., McKenzie, L.M., Hays, S.M., 2012. Interpreting variability in population biomonitoring data: role of elimination kinetics. J. Expo. Sci. Environ. Epidemiol. 22, 398–408. https://doi.org/10.1038/jes.2012.35.

Baas, J., Schotten, M., Plume, A., Côté, G., Karimi, R., 2020. Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. Quant. Sci. Stud. 1, 377–386. https://doi.org/10.1162/qss_a_00019.

Barnett-Itzhaki, Z., Esteban López, M., Puttaswamy, N., Berman, T., 2018. A review of human biomonitoring in selected Southeast Asian countries. Environ. Int. 116, 156–164. https://doi.org/10.1016/j.envint.2018.03.046.

Barr, D.B., Wilder, L.C., Caudill, S.P., Gonzalez, A.J., Needham, L.L., Pirkle, J.L., 2005. Urinary creatinine concentrations in the U.S. population: Implications for urinary biologic monitoring measurements. Environ. Health Perspect. 113, 192–200. https://doi.org/10.1289/ehp.7337.

Basu, N., Horvat, M., Evers, D.C., Zastenskaya, I., Weihe, P., Tempowski, J., 2018. A state-of-the-science review of mercury biomarkers in human populations worldwide between 2000 and 2018. Environ. Health Perspect. 126 https://doi.org/10.1289/EHP3904.

Blettner, M., Sauerbrei, W., Schlehofer, B., Scheuchenpflug, T., Friedenreich, C., 1999. Traditional reviews, meta-analyses and pooled analyses in epidemiology. Int. J. Epidemiol. 28, 1–9. https://doi.org/10.1093/ije/28.1.1.

Falagas, M.E., Pitsouni, E.I., Malietzis, G.A., Pappas, G., 2008. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. FASEB J. 22, 338–342. https://doi.org/10.1096/fj.07-9492LSF.

Farzanfar, D., Abumuamar, A., Kim, J., Sirotich, E., Wang, Y., Pullenayegum, E., 2017. Longitudinal studies that use data collected as part of usual care risk reporting biased results: a systematic review. BMC Med. Res. Methodol. 17, 133. https://doi.org/10.1186/s12874-017-0418-1.

Finkelstein, M.M., Verma, D.K., 2001. Exposure estimation in the presence of nondetectable values: Another look. Am. Ind. Hyg. Assoc. J. 62, 195–198. https://doi.org/10.1080/15298660108984622.

Fisher, R.A., 1921. 014: On the "Probable Error" of a Coefficient of Correlation Deduced from a Small Sample. Metron 1, 3–32.

Fisher, R.A., 1915. Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. Biometrika 10, 507. https://doi.org/10.2307/2331838.

Frandsen, T.F., Eriksen, M.B., Hammer, D.M.G., Christensen, J.B., 2019. PubMed coverage varied across specialties and over time: a large-scale study of included studies in Cochrane reviews. J. Clin. Epidemiol. 112, 59–66. https://doi.org/10.1016/j.jclinepi.2019.04.015.

Frank, J.J., Poulakos, A.G., Tornero-Velez, R., Xue, J., 2019. Systematic review and meta-analyses of lead (Pb) concentrations in environmental media (soil, dust, water, food, and air) reported in the United States from 1996 to 2016. Sci. Total Environ. https://doi.org/10.1016/J.SCITOTENV.2019.07.295.

Gauss–Markov Theorem, 2008. In: The Concise Encyclopedia of Statistics. Springer, New York, pp. 217–218. https://doi.org/10.1007/978-0-387-32833-1_159.

Gilliom, R.J., Hirsch, R.M., Gilroy, E.J., 1984. Effect of censoring trace-level water-quality data on trend-detection capability. Environ. Sci. Technol. 18, 530–535. https://doi.org/10.1021/es00125a009.

Godin, K., Stapleton, J., Kirkpatrick, S.I., Hanning, R.M., Leatherdale, S.T., 2015. Applying systematic review search methods to the grey literature: A case study examining guidelines for school-based breakfast programs in Canada. Syst. Rev. 4, 138. https://doi.org/10.1186/s13643-015-0125-0.

Grant, K., Goldizen, F.C., Sly, P.D., Brune, M.-N., Neira, M., van den Berg, M., Norman, R. E., 2013. Health consequences of exposure to e-waste: a systematic review. Lancet. Glob. Heal. 1, e350–e361. https://doi.org/10.1016/S2214-109X(13)70101-3.

Gusenbauer, M., Haddaway, N.R., 2020a. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. Res. Synth. Methods 11, 181–217. https://doi.org/10.1002/jrsm.1378.

Gusenbauer, M., Haddaway, N.R., 2020b. What every Researcher should know about Searching – Clarified Concepts, Search Advice, and an Agenda to improve Finding in Academia. Res. Synth. Methods jrsm.1457. https://doi.org/10.1002/jrsm.1457.

Guyatt, G.H., Oxman, A.D., Vist, G.E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., Schünemann, H.J., 2008. GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. BMJ 336, 924–926. https://doi.org/10.1136/bmj.39489.470347.ad.

Haddaway, N.R., Bernes, C., Jonsson, B.-G., Hedlund, K., 2016. The benefits of systematic mapping to evidence-based environmental management. Ambio 45, 613–620. https://doi.org/10.1007/s13280-016-0773-x.

Haddaway, N.R., Macura, B., Whaley, P., Pullin, A.S., 2018. ROSES Reporting standards for Systematic Evidence Syntheses: Pro forma, flow-diagram and descriptive summary of the plan and conduct of environmental systematic reviews and systematic maps. Environ. Evid. 7, 7. https://doi.org/10.1186/s13750-018-0121-7.

Halperin, M., 1960. Extension of the Wilcoxon-Mann-Whitney Test to Samples Censored at the Same Fixed Point. J. Am. Stat. Assoc. 55, 125. https://doi.org/10.2307/2282183.

Harari, M.B., Parola, H.R., Hartwell, C.J., Riegelman, A., 2017. Literature searches in systematic reviews and meta-analyses: a review, evaluation, and recommendations. J. Vocat. Behav. 118 https://doi.org/10.1016/j.jvb.2020.103377.

Heffernan, A.L., Aylward, L.L., Toms, L.-M.L., Sly, P.D., Macleod, M., Mueller, J.F., 2014. Pooled biological specimens for human biomonitoring of environmental chemicals: opportunities and limitations. J. Expo. Sci. Environ. Epidemiol. 24, 225–232. https://doi.org/10.1038/jes.2013.76.

Hoffmann, S., de Vries, R.B.M., Stephens, M.L., Beck, N.B., Dirven, H.A.A.M., Fowle, J.R., Goodman, J.E., Hartung, T., Kimber, I., Lalu, M.M., Thayer, K., Whaley, P., Wikoff, D., Tsaioun, K., 2017. A primer on systematic reviews in toxicology. Arch. Toxicol. 91, 2551–2575. https://doi.org/10.1007/s00204-017-1980-3.

Hornung, R.W., Reed, L.D., 1990. Estimation of average concentration in the presence of nondetectable values. Appl. Occup. Environ. Hyg. 5, 46–51. https://doi.org/10.1080/1047322X.1990.10389587.

Huth, R., 1999. Testing for trends in data unevenly distributed in time. Theor. Appl. Climatol. 64, 151–162. https://doi.org/10.1007/s007040050119.

Kohl, C., McIntosh, E.J., Unger, S., Haddaway, N.R., Kecke, S., Schiemann, J., Wilhelm, R., 2018. Online tools supporting the conduct and reporting of systematic reviews and systematic maps: a case study on CADIMA and review of existing tools. Environ. Evid. 7, 1–17. https://doi.org/10.1186/s13750-018-0115-5.

Kolmogorov, A.N., 1933. Sulla Determinazione Empirica di Una Legge di Distribuzione. G. dell'Istituto Ital. degli Attuari 4, 83–91.

Kruskal, W.H., Wallis, W.A., 1952. Use of Ranks in One-Criterion Variance Analysis. J. Am. Stat. Assoc. 47, 583–621. https://doi.org/10.1080/01621459.1952.10483441.

Larsen, P.O., von Ins, M., 2010. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. Scientometrics 84, 575–603. https://doi.org/10.1007/s11192-010-0202-z.

Larsen, P.S., Kamper-Jørgensen, M., Adamson, A., Barros, H., Bonde, J.P., Brescianini, S., Brophy, S., Casas, M., Devereux, G., Eggesbø, M., Fantini, M.P., Frey, U., Gehring, U., Grazuleviciene, R., Henriksen, T.B., Hertz-Picciotto, I., Heude, B., Hryhorczuk, D.O., Inskip, H., Jaddoe, V.W.V., Lawlor, D.A., Ludvigsson, J., Kelleher, C., Kiess, W., Koletzko, B., Kuehni, C.E., Kull, I., Kyhl, H.B., Magnus, P., Momas, I., Murray, D., Pekkanen, J., Polanska, K., Porta, D., Poulsen, G., Richiardi, L., Roeleveld, N., Skovgaard, A.M., Sram, R.J., Strandberg-Larsen, K., Thijs, C., Van Eijsden, M., Wright, J., Vrijheid, M., Andersen, A.-M.N., 2013. Pregnancy and Birth Cohort Resources in Europe: a Large Opportunity for Aetiological Child Health Research. Paediatr. Perinat. Epidemiol. 27, 393–414. https://doi.org/10.1111/ppe.12060.

Leenaars, M., Hooijmans, C.R., van Veggel, N., ter Riet, G., Leeflang, M., Hooft, L., van der Wilt, G.J., Tillema, A., Ritskes-Hoitinga, M., 2012. A step-by-step guide to

systematically identify all relevant animal studies. Lab. Anim. 46, 24–31. https://doi.org/10.1258/la.2011.011087.
Lermen, D., Bartel-Steinbach, M., Gwinner, F., Conrad, A., Weber, T., von Briesen, H., Kolossa-Gehring, M., 2019. Trends in characteristics of 24-h urine samples and their relevance for human biomonitoring studies – 20 years of experience in the German Environmental Specimen Bank. Int. J. Hyg. Environ. Health 222, 831–839. https://doi.org/10.1016/j.ijheh.2019.04.009.
Lewis, K.M., Ruiz, M., Goldblatt, P., Morrison, J., Porta, D., Forastiere, F., Hryhorczuk, D., Zvinchuk, O., Saurel-Cubizolles, M.J., Lioret, S., Annesi-Maesano, I., Vrijheid, M., Torrent, M., Iniguez, C., Larranaga, I., Harskamp-van Ginkel, M.W., Vrijkotte, T.G. M., Klanova, J., Svancara, J., Barross, H., Correia, S., Jarvelin, M.R., Taanila, A., Ludvigsson, J., Faresjo, T., Marmot, M., Pikhart, H., 2017. Mother's education and offspring asthma risk in 10 European cohort studies. Eur. J. Epidemiol. 32, 797–805. https://doi.org/10.1007/s10654-017-0309-0.
Liberati, A., Altman, D.G., Tetzlaff, J., Mulrow, C., Gøtzsche, P.C., Ioannidis, J.P.A., Clarke, M., Devereaux, P.J., Kleijnen, J., Moher, D., 2009. PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 339. https://doi.org/10.1136/bmj.b2700.
Limpert, E., Stahel, W.A., Abbt, M., 2001. Log-normal distributions across the sciences: keys and clues. Bioscience 51, 341–352. https://doi.org/10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2.
Mann, H.B., Whitney, D.R., 1947. On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Stat. 18, 50–60. https://doi.org/10.1214/aoms/1177730491.
Mary-Huard, T., Daudin, J.-J., Baccini, M., Biggeri, A., Bar-Hen, A., 2007. Biases induced by pooling samples in microarray experiments. Bioinformatics 23, i313–i318. https://doi.org/10.1093/bioinformatics/btm182.
Matoušek, J., Mount, D.M., Netanyahu, N., 1998. Efficient randomized algorithms for the repeated median line estimator. Algorithmica 20, 136–150.
Morgan, R.L., Beverly, B., Ghersi, D., Schünemann, H.J., Rooney, A.A., Whaley, P., Zhu, Y.G., Thayer, K.A., 2019. GRADE guidelines for environmental and occupational health: A new series of articles in Environment International. Environ. Int. https://doi.org/10.1016/j.envint.2019.04.016.
Morgan, R.L., Thayer, K.A., Bero, L., Bruce, N., Falck-Ytter, Y., Ghersi, D., Guyatt, G., Hooijmans, C., Langendam, M., Mandrioli, D., Mustafa, R.A., Rehfuess, E.A., Rooney, A.A., Shea, B., Silbergeld, E.K., Sutton, P., Wolfe, M.S., Woodruff, T.J., Verbeek, J.H., Holloway, A.C., Santesso, N., Schünemann, H.J., 2016. GRADE: Assessing the quality of evidence in environmental and occupational health. Environ. Int. 92–93, 611–616. https://doi.org/10.1016/j.envint.2016.01.004.
Morgan, R.L., Whaley, P., Thayer, K.A., Schünemann, H.J., 2018. Identifying the PECO: A framework for formulating good questions to explore the association of environmental and other exposures with health outcomes. Environ. Int. https://doi.org/10.1016/j.envint.2018.07.015.
Onofri, A., Piepho, H.P., Kozak, M., 2019. Analysing censored data in agricultural research: A review with examples and software tips. Ann. Appl. Biol. https://doi.org/10.1111/aab.12477.
Page, M.J., Moher, D., Bossuyt, P., Boutron, I., Hoffmann, T., Mulrow, C., Shamseer, L., Tetzlaff, J., Akl, E., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J., Hróbjartsson, A., Lalu, M., Li, T., Loder, E., Mayo-Wilson, E., McDonald, S., McGuinness, L.A., Stewart, L., Thomas, J., Tricco, A., Welch, V., Whiting, P., McKenzie, J., 2020. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. MetaArXiv. https://doi.org/10.31222/OSF.IO/GWDHK.
Patel, C.J., Pho, N., McDuffie, M., Easton-Marks, J., Kothari, C., Kohane, I.S., Avillach, P., 2016. A database of human exposomes and phenomes from the US National Health and Nutrition Examination Survey. Sci. Data 3, 1–10. https://doi.org/10.1038/sdata.2016.96.
Rethlefsen, M.L., Farrell, A.M., Osterhaus Trzasko, L.C., Brigham, T.J., 2015. Librarian co-authors correlated with higher quality reported search strategies in general internal medicine systematic reviews. J. Clin. Epidemiol. 68, 617–626. https://doi.org/10.1016/j.jclinepi.2014.11.025.
Rethlefsen, M.L., Farrell, A.M., Osterhaus Trzasko, L.C., Brigham, T.J., 2015. Librarian co-authors correlated with higher quality reported search strategies in general internal medicine systematic reviews. J. Clin. Epidemiol. 68, 617–626. https://doi.org/10.1016/j.jclinepi.2014.11.025.
Schulz, C., Angerer, J., Ewers, U., Kolossa-Gehring, M., 2007. The german human biomonitoring commission. Int. J. Hyg. Environ. Health 210, 373–382. https://doi.org/10.1016/j.ijheh.2007.01.035.
Schünemann, H.J., Moja, L., 2015. Reviews: Rapid! Rapid! Rapid!.and systematic. Syst. Rev. https://doi.org/10.1186/2046-4053-4-4.
Sen, P.K., 1968. Estimates of the Regression Coefficient Based on Kendall's Tau. J. Am. Stat. Assoc. 63, 1379–1389. https://doi.org/10.1080/01621459.1968.10480934.
Shapiro, S.S., Wilk, M.B., 1965. An analysis of variance test for normality (complete samples). Biometrika 52, 591–611. https://doi.org/10.1093/biomet/52.3-4.591.
Sharma, B.M., Bharat, G.K., Tayal, S., Nizzetto, L., Čupr, P., Larssen, T., 2014. Environment and human exposure to persistent organic pollutants (POPs) in India: a systematic review of recent and historical data. Environ. Int. 66, 48–64. https://doi.org/10.1016/j.envint.2014.01.022.
Sharma, B.M., Sáňka, O., Kalina, J., Scheringer, M., 2019. An overview of worldwide and regional time trends in total mercury levels in human blood and breast milk from 1966 to 2015 and their associations with health effects. Environ. Int. 125, 300–319. https://doi.org/10.1016/J.ENVINT.2018.12.016.
Shih, W.J., Binkowitz, B., 1987. C282. Median versus geometric mean for lognormal samples. J. Stat. Comput. Simul. 28, 81–83. https://doi.org/10.1080/00949658708811013.
Simpson, E.H., 1951. The Interpretation of Interaction in Contingency Tables, Source: Journal of the Royal Statistical Society. Series B (Methodological).
Smirnov, N., 1948. Table for estimating the goodness of fit of empirical distributions. Ann. Math. Stat. 19, 279–281. https://doi.org/10.1214/aoms/1177730256.
Student, 1908. The probable error of a mean. Biometrika 6, 1–25. https://doi.org/10.1093/biomet/6.1.1.
Theil, H., 1992. A Rank-Invariant Method of Linear and Polynomial Regression Analysis. Springer, Dordrecht, pp. 345–381. https://doi.org/10.1007/978-94-011-2546-8_20.
Tober, M., 2011. PubMed, ScienceDirect, Scopus or Google Scholar – Which is the best search engine for an effective literature research in laser medicine? Med. Laser Appl. 26, 139–144. https://doi.org/10.1016/j.mla.2011.05.006.
Uman, L.S., 2011. Systematic reviews and meta-analyses. J. Can. Acad. Child Adolesc. Psychiatry 20, 57–59.
University of Michigan Library, 2020. Research Guides: Systematic Reviews: Creating a Search Strategy.
Vandenberg, L.N., Ågerstrand, M., Beronius, A., Beausoleil, C., Bergman, Å., Bero, L.A., Bornehag, C.-G., Boyer, C.S., Cooper, G.S., Cotgreave, I., Gee, D., Grandjean, P., Guyton, K.Z., Hass, U., Heindel, J.J., Jobling, S., Kidd, K.A., Kortenkamp, A., Macleod, M.R., Martin, O. V., Norinder, U., Scheringer, M., Thayer, K.A., Toppari, J., Whaley, P., Woodruff, T.J., Rudén, C., 2016. A proposed framework for the systematic review and integrated assessment (SYRINA) of endocrine disrupting chemicals. Environ. Heal. 15, 74. https://doi.org/10.1186/s12940-016-0156-6.
Wang, Z., Nayfeh, T., Tetzlaff, J., O'Blenis, P., Murad, M.H., 2020. Error rates of human reviewers during abstract screening in systematic reviews. PLoS One 15. https://doi.org/10.1371/journal.pone.0227742.
Weiss, F.T., Leuzinger, M., Zurbrugg, C., Eggen, R.I.L., 2016. Chemical Pollution in Low- and Middle-Income Countries. Dubendorf.
Whaley, P., Aiassa, E., Beausoleil, C., Beronius, A., Bilotta, G., Boobis, A., de Vries, R., Hanberg, A., Hoffmann, S., Hunt, N., Kwiatkowski, C.F., Lam, J., Lipworth, S., Martin, O., Randall, N., Rhomberg, L., Rooney, A.A., Schünemann, H.J., Wikoff, D., Wolffe, T., Halsall, C., 2020a. Recommendations for the conduct of systematic reviews in toxicology and environmental health research (COSTER). Environ. Int. 143, 105926. https://doi.org/10.1016/j.envint.2020.105926.
Whaley, P., Edwards, S.W., Kraft, A., Nyhan, K., Shapiro, A., Watford, S., Wattam, S., Wolffe, T., Angrish, M., 2020b. Knowledge organization systems for systematic chemical assessments. Environ. Health Perspect. 128 https://doi.org/10.1289/EHP6994.
Whaley, P., Halsall, C., Ågerstrand, M., Aiassa, E., Benford, D., Bilotta, G., Coggon, D., Collins, C., Dempsey, C., Duarte-Davidson, R., FitzGerald, R., Galay-Burgos, M., Gee, D., Hoffmann, S., Lam, J., Lasserson, T., Levy, L., Lipworth, S., Ross, S.M., Martin, O., Meads, C., Meyer-Baron, M., Miller, J., Pease, C., Rooney, A., Sapiets, A., Stewart, G., Taylor, D., 2016. Implementing systematic review techniques in chemical risk assessment: Challenges, opportunities and recommendations. Environ. Int. 92–93, 556–564. https://doi.org/10.1016/j.envint.2015.11.002.
WHO, 2015. Human biomonitoring: facts and figures. Copenhagen.
Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biometrics Bull. 1, 80. https://doi.org/10.2307/3001968.
Wolffe, T.A.M., Whaley, P., Halsall, C., Rooney, A.A., Walker, V.R., 2019. Systematic evidence maps as a novel tool to support evidence-based decision-making in chemicals policy and risk management. Environ. Int. 130 https://doi.org/10.1016/j.envint.2019.05.065.
Zhang, X., Zwiers, F.W., Li, G., 2004. Monte Carlo experiments on the detection of trends in extreme values. J. Clim. 17, 1945–1952 https://doi.org/10.1175/1520-0442(2004)017<1945:MCEOTD>2.0.CO;2.