United States
Environmental Protection Agency

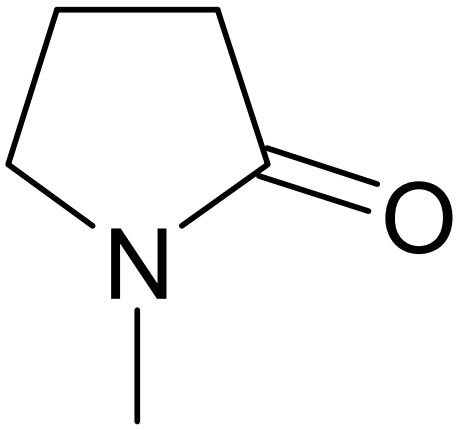
Office of Chemical Safety and
Pollution Prevention

Final Risk Evaluation for n-Methylpyrrolidone

Systematic Review Supplemental File:

Updates to the Data Quality Criteria for Epidemiological Studies

CASRN: 872-50-4



December 2020

EPA's Office of Pollution Prevention and Toxics (OPPT) developed data quality criteria for epidemiological studies. The first version of the criteria was documented in the *Application of Systematic Review in TSCA Risk Evaluations* document (EPA Document #740-P1-8001). The initial criteria were updated after considering EPA/OPPT's practical experience and comments from the public. This systematic review supplemental document describes the updated data quality criteria for epidemiological studies that EPA/OPPT intends to apply for the TSCA risk evaluations. Refer to Appendix H of the *Application of Systematic Review in TSCA Risk Evaluations* document for details about the data quality evaluation tool.

Evaluation Criteria for Epidemiological Studies: General

Confidence Level (Score)	Description	Selected Score
Domain 1. Study Participation		
Metric 1. Participant selection (selection, performance biases)		
Instructions: To meet criteria for confidence ratings for metrics where 'AND' is included, studies must address both conditions where "AND" is stipulated. To meet criteria for confidence ratings for metrics where 'OR' is included studies must address at least one of the conditions stipulated.		
High (score = 1)	• ***For all study types:*** All key elements of the study design are reported (e.g., setting, participation rate described at all steps of the study, inclusion and exclusion criteria, and methods of participant selection or case ascertainment) **AND** The reported information indicates that selection in or out of the study (or analysis sample) and participation was not likely to be biased (i.e., the exposure-outcome distribution of the participants is likely representative of the exposure-outcome distributions in the population of persons eligible for inclusion in the study.)	
Medium (score = 2)	• ***For all study types:*** Some key elements of the study design were not present but available information indicates a low risk of selection bias (i.e., the exposure-outcome distribution of the participants is likely representative of the exposure-outcome distributions in the population of persons eligible for inclusion in the study.)	
Low (score = 3)	• ***For all study types:*** Key elements of the study design and information on the population (e.g., setting, participation rate described at most steps of the study, inclusion and exclusion criteria, and methods of participant selection or case ascertainment) are not reported [STROBE checklist 4, 5 and 6 (Von Elm et al., 2008)].	
Unacceptable (score = 4)	***For all study types:*** The reported information indicates that selection in or out of the study (or analysis sample) and participation was likely to be significantly biased (i.e., the exposure-outcome distribution of the participants is likely not representative of the exposure-outcome distribution of the population of persons eligible for inclusion in the study.)	
Not rated/ not applicable (NA)	Do not select for this metric.	
Reviewer's comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*	

Confidence Level (Score)	Description	Selected Score
Metric 2. Attrition (missing data/attrition/exclusion, reporting biases)		
High (score = 1)	• ***For cohort studies:*** There was minimal subject loss to follow up during the study (or exclusion from the analysis sample) and outcome and exposure data were largely complete. **OR** • Any loss of subjects (i.e., incomplete outcome data) or missing exposure and outcome data were adequately* addressed (as described below) and reasons were documented when human subjects were removed from a study (NTP, 2015). **AND** • Missing data have been imputed using appropriate methods (e.g., multiple imputation methods), and characteristics of subjects lost to follow up or with unavailable records are not significantly different from those of the study participants (NTP, 2015). • ***For case-control studies and cross-sectional studies:*** There was minimal subject withdrawal from the study (or exclusion from the analysis sample) and outcome data and exposure were largely complete. **OR** • Any exclusion of subjects from analyses was adequately* addressed (as described below), and reasons were documented when subjects were removed from the study or excluded from analyses (NTP, 2015). ***NOTE for all study types:*** Adequate handling of subject attrition can include: Use of imputation methods for missing outcome and exposure data; reasons for missing subjects unlikely to be related to outcome (for survival data, censoring was unlikely to introduce bias); missing outcome data balanced in numbers across study groups, with similar reasons for missing data across groups.	
Medium (score = 2)	• ***For cohort studies:*** There was moderate subject loss to follow up during the study (or exclusion from the analysis sample) or outcome and exposure data were nearly complete. **AND** • Any loss or exclusion of subjects was adequately addressed (as described in the acceptable handling of subject attrition in the high confidence category) and reasons were documented when human subjects were removed from a study. • ***For case-control studies and cross-sectional studies:*** There was moderate subject withdrawal from the study (or exclusion from the analysis sample), but outcome and exposure data were largely complete **AND** • Any exclusion of subjects from analyses was adequately addressed (as described above), and reasons were documented when subjects were removed from the study or excluded from analyses (NTP, 2015).	
Low (score = 3)	***For cohort studies:*** The loss of subjects (e.g., loss to follow up, incomplete outcome or exposure data) was moderate and unacceptably handled (as described below in the unacceptable confidence category) (NTP, 2015). **OR** • Numbers of individuals were not reported at important stages of study (e.g., numbers of eligible participants included in the study or analysis sample, completing follow-up, and analyzed). Reasons were not provided for non-participation at each stage (Von Elm et al., 2008).	

Confidence Level (Score)	Description	Selected Score
	For case-control and cross-sectional studies: The exclusion of subjects from analyses was moderate and unacceptably handled (as described below in the unacceptable confidence category). **OR** • Numbers of individuals were not reported at important stages of study (e.g., numbers of eligible participants included in the study or analysis sample, completing follow-up, and analyzed). Reasons were not provided for non-participation at each stage (Von Elm et al., 2008).	
Unacceptable (score = 4)	• ***For cohort studies:*** There was large subject attrition during the study (or exclusion from the analysis sample). **OR** • Unacceptable handling of subject attrition: reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across study groups; or potentially inappropriate application of imputation (NTP, 2015). • ***For case-control and cross-sectional studies:*** There was large subject withdrawal from the study (or exclusion from the analysis sample). **OR** • Unacceptable handling of subject attrition: reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across study groups; or potentially inappropriate application of imputation.	
Not rated/applicable	• Do not select for this metric.	
Reviewer's comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*	
Metric 3. Comparison Group (selection, performance biases)		
High (score = 1)	• ***For ALL study types:*** Any differences in baseline characteristics of groups were considered as potential confounding or stratification variables and were thereby controlled by statistical analysis (NTP, 2015). **OR** • ***For cohort and cross-sectional studies:*** Key elements of the study design are reported (i.e., setting, inclusion and exclusion criteria, and methods of participant selection), and indicate that subjects were similar (e.g., recruited from the same eligible population with the same method of ascertainment and within the same time frame using the same inclusion and exclusion criteria, and were of similar age and health status) (NTP, 2015). • ***For case-control studies:*** Key elements of the study design are reported indicate that that cases and controls were similar (e.g., recruited from the same eligible population with the number of controls described, and eligibility criteria and are recruited within the same time frame (NTP, 2015). • ***For studies reporting Standardized Mortality Ratios (SMRs) or Standardized Incidence Ratios (SIRs):*** Age, sex (if applicable), and race (if applicable) adjustment or stratification is described and choice of reference population (e.g., general population) is reported.	

Confidence Level (Score)	Description	Selected Score
Medium (score = 2)	• ***For cohort studies and cross-sectional studies:*** There is only indirect evidence (e.g., stated by the authors without providing a description of methods) that groups are similar (as described above for the high confidence rating). • ***For case-control studies***: There is indirect evidence (i.e., stated by the authors without providing a description of methods) that cases and controls are similar (as described above for the high confidence rating). • ***For studies reporting SMRs or SIRs:*** Age, sex (if applicable), and race (if applicable) adjustment or stratification is not specifically described in the text, but results tables are stratified by age and/or sex (i.e., indirect evidence); choice of reference population (e.g., general population) is reported.	
Low (score = 3)	• ***For cohort and cross-sectional studies***: There is indirect evidence (i.e., stated by the authors without providing a description of methods) that groups were not similar (as described above for the high confidence rating). **AND** • Control for differences in exposure groups is not adequately controlled for in the statistical analysis. • ***For case-control studies***: There is indirect evidence (i.e., stated by the authors without providing a description of methods) that cases and controls were not similar (as described above for the high confidence rating). **AND** • The characteristics of cases and controls are not reported (NTP, 2015). **AND** • Control for differences in the case and control groups is not adequately controlled for in the statistical analysis. • ***For studies reporting SMRs or SIRs***: Indirect evidence of a lack of adjustment or stratification for age or sex (if applicable); indirect evidence that choice of reference population (e.g., general population) is appropriate.	
Unacceptable (score = 4)	• ***For cohort studies:*** Subjects in all exposure groups were not similar **OR** • Information was not reported to determine if participants in all exposure groups were similar [STROBE Checklist 6 (Von Elm et al., 2008) **AND** • Potential differences in exposure groups were not controlled for in the statistical analysis. **OR** • Subjects in the exposure groups had very different participation/response rates (NTP, 2015). • ***For case-control studies:*** Controls were drawn from a very dissimilar population than cases or recruited within very different time frames (NTP, 2015). **AND** • Potential differences in the case and control groups were not controlled for in the statistical analysis. **OR**	

Confidence Level (Score)	Description	Selected Score
	• Rationale and/or methods for case and control selection, matching criteria including number of controls per case (if relevant) were not reported [STROBE Checklist 6 (Von Elm et al., 2008)]. • ***For cross-sectional studies:*** Subjects in all exposure groups were not similar, recruited within very different time frames, or had very different participation/response rates (NTP, 2015). **AND** • Potential differences in exposure groups were not controlled for in the statistical analysis. **OR** • Sources and methods of selection of participants in all exposure groups were not reported [STROBE Checklist 6 (Von Elm et al., 2008)]. • ***For studies reporting SMRs or SIRs:*** Lack of adjustment or stratification for both age and sex (if applicable); choice of reference population (e.g., general population) is not reported.	
Not rated/applicable	• Do not select for this metric.	
Reviewer's comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*	
Domain 2. Exposure Characterization		
Metric 4. Measurement of Exposure (Detection/measurement/information, performance biases)		
High (score = 1)	• ***For all study types:*** Exposure was consistently assessed (i.e., using the same method and sampling time-frame) using well-established methods (e.g., personal and/or industrial hygiene data used to determine levels of exposure, a frequently used biomarker of exposure) that directly measure exposure [e.g., measurement of the chemical in the environment (air, drinking water, consumer product] or measurement of the chemical concentration in a biological matrix (e.g., blood, plasma, urine) (NTP, 2015). **OR** • For an occupational population, contains detailed employment records which allows for construction of a job-matrix for entire work history of exposure (i.e., cumulative or peak exposures, and time since first exposure).	
Medium (score = 2)	• ***For all study types:*** Exposure was directly measured and assessed using a method that is not well-established (e.g., newly developed biomarker of exposure), ***but*** is validated against a well-established method and demonstrated a high agreement between the two methods **OR** • For an occupational study population, contains detailed employment records for only a portion of participant's work history. (i.e., only early years or later years), such that extrapolation of the missing years is required.	
Low (score = 3)	• ***For all study types:*** A less-established method (e.g., newly developed biomarker of exposure) was used and no method validation was conducted against well-established methods, but there was little to no evidence that the method had poor validity and little to no evidence of significant exposure misclassification (e.g., differential recall of self-reported exposure) (NTP, 2015). **OR**	

Confidence Level (Score)	Description	Selected Score
	• For an occupational study population, exposure was estimated solely using professional judgement.	
Unacceptable (score = 4)	• ***For all study types:*** Methods used to quantify the exposure were not well defined, and sources of data and detailed methods of exposure assessment were not reported [STROBE Checklist 7 and 8] **OR** • Exposure was assessed using methods known or suspected to have poor validity (NTP, 2015). **OR** • There is evidence of substantial exposure misclassification that would significantly bias the results.	
Not rated/applicable	• Do not select for this metric.	
Reviewer's comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*	
Metric 5. Exposure levels (Detection/measurement/information biases)		
High (score = 1)	• Do not select for this metric.	
Medium (score = 2)	• ***For all study types:*** The range and distribution of exposure is sufficient or adequate to develop an exposure-response estimate (Cooper et al., 2016). **AND** • Reports 3 or more levels of exposure (i.e., referent group and 2 or more) or an exposure-response model using a continuous measure of exposure.	
Low (score = 3)	• ***For all study types:*** The range of exposure in the population is limited **OR** • Reports 2 levels of exposure (e.g., exposed/unexposed)) (Cooper et al., 2016)	
Unacceptable (score = 4)	• ***For all study types:*** The range and distribution of exposure are not adequate to determine an exposure-response relationship (Cooper et al., 2016). **OR** • No description is provided on the levels or range of exposure.	
Not rated/applicable	• Do not select for this metric.	
Reviewer's comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*	
Metric 6. Temporality (Detection/measurement/information biases)		
High (score = 1)	• ***For all study types:*** The study presents an appropriate temporality between exposure and outcome (i.e., the exposure precedes the disease). **AND** • The interval between the exposure (or reconstructed exposure) and the outcome has an appropriate consideration of relevant exposure windows (Lakind et al., 2014).	
Medium (score = 2)	• ***For all study types:*** Temporality is established, but it is unclear whether exposures fall within relevant exposure windows for the outcome of interest (Lakind et al., 2014).	
Low (score = 3)	• ***For all study types:*** The temporality of exposure and outcome is uncertain	

Confidence Level (Score)	Description	Selected Score
Unacceptable (score = 4)	• ***For all study types:*** Study lacks an established time order, such that exposure is not likely to have occurred prior to outcome (Lakind et al., 2014). **OR** • There was inadequate follow-up of the cohort for the expected latency period. **OR** • Sources of data and details of methods of assessment were not sufficiently reported (e.g., duration of follow-up, periods of exposure, dates of outcome ascertainment) [STROBE Checklist 8 (Von Elm et al., 2008)].	
Not rated/applicable	• Do not select for this metric.	
Reviewer's comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*	
Domain 3. Outcome Assessment		
Metric 7. Outcome measurement or characterization (detection/measurement/information, performance, reporting biases)		
High (score = 1)	• ***For cohort studies:*** The outcome was assessed using well-established methods (e.g., the "gold standard"). ***For case-control studies:*** The outcome was assessed in cases (i.e., case definition) and controls using well-established methods (the gold standard). Subjects had been followed for the same length of time in all study groups (NTP, 2015). • ***For cross-sectional studies***: There is direct evidence that the outcome was assessed using well-established methods (the gold standard) (NTP, 2015). ***Note**: Acceptable assessment methods will depend on the outcome, but examples of such methods may include: objectively measured with diagnostic methods, measured by trained interviewers, obtained from registries (NTP, 2015; Shamliyan et al., 2010).	
Medium (score = 2)	• ***For all study types:*** A less-established method was used and no method validation was conducted against well-established methods, but there was little to no evidence that that the method had poor validity and little to no evidence of outcome misclassification (e.g., differential reporting of outcome by exposure status).	
Low (score = 3)	• ***For cohort studies:*** The outcome assessment method is an insensitive instrument or measure. **OR** • The length of follow up differed by study group (NTP, 2015). • ***For case-control studies:*** The outcome was assessed in cases (i.e., case definition) using an insensitive instrument or measure (NTP, 2015). • ***For cross-sectional studies:*** The outcome assessment method is an insensitive instrument or measure (NTP, 2015). • Any self-reported information	
Unacceptable (score = 4)	• ***For all study types:*** Diagnostic criteria were not defined or reported [STROBE Checklist 15 (Von Elm et al., 2008)].	
Not rated/applicable	• Do not select for this metric	

Confidence Level (Score)	Description	Selected Score
Reviewer's comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*	
Metric 8. Reporting Bias		
High (score = 1)	• ***For all study types:*** A description of measured outcomes is reported in the methods, abstract, and/or introduction. Effect estimates are reported with a confidence interval and/or standard errors; number of cases/controls or exposed/unexposed reported for each analysis, to be included in exposure-response analysis or fully tabulated during data extraction and analyses (NTP, 2015).	
Medium (score = 2)	• ***For all study types:*** All of the study's measured outcomes (primary and secondary) outlined in the methods, abstract, and/or introduction (that are relevant for the evaluation) are reported, but not in a way that would allow for detailed extraction (e.g., results were discussed in the text but accompanying data were not shown).	
Low (score = 3)	• ***For all study types:*** All of the study's measured outcomes (primary and secondary) outlined in the methods, abstract, and/or introduction (that are relevant for the evaluation) have not been reported. ***Note:*** In addition to not reporting outcomes, this would include reporting outcomes based on composite score without individual outcome components or outcomes reported using measurements, analysis methods, or unplanned analyses were included that would appreciably bias results (NTP, 2015).	
Unacceptable (score = 4)	• Do not select for this metric.	
Not rated/applicable	• Do not select for this metric.	
Reviewer's comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*	
Domain 4. Potential Confounding/Variable Control		
Metric 9. Covariate Adjustment (confounding)		
High (score = 1)	• ***For all study types:*** Appropriate adjustments or explicit considerations were made for potential confounders (e.g., age, sex, socioeconomic status) (excluding co-exposures, which are evaluated in metric 11) in the final analyses through the use of statistical models to reduce research-specific bias, including matching, adjustment in multivariate models, stratification, or other methods that were appropriately justified (NTP, 2015). • ***For Studies reporting SMRs or SIRs:*** Adjustments are described and results are age-, race-, and sex-adjusted (or stratified) if applicable..	
Medium (score = 2)	• ***For all study types:*** There is indirect evidence that appropriate adjustments were made [i.e., considerations were made for potential confounders (excluding co-exposures)] without providing a description of methods. **OR** • The distribution of potential confounders (excluding co-exposures) did not differ significantly between exposure groups or between cases and controls. **OR** • The major potential confounders (excluding co-exposures) were appropriately adjusted (e.g., SMRs, SIRs) and any not adjusted for are considered not to appreciably bias the results	

Confidence Level (Score)	Description	Selected Score
	• ***For Studies reporting SMRs or SIRs:*** Indirect evidence that results are age- and sex-adjusted (or stratified) if applicable.	
Low (score = 3)	• ***For all study types:*** There is indirect evidence (i.e., no description is provided in the study) that considerations were not made for potential confounders adjustment in the final analyses (NTP, 2015). **AND** • The distribution of primary covariates (excluding co-exposures) and potential confounders was not reported between the exposure groups or between cases and controls (NTP, 2015). • ***For Studies reporting SMRs or SIRs:*** Results are age-, race-, OR sex-adjusted (or stratified) if applicable (i.e., if both *should* have been adjusted).	
Unacceptable (score = 4)	• ***For all study types:*** The distribution of potential confounders differed significantly between the exposure group. **AND** • Confounding was demonstrated and was not appropriately adjusted for in the final analyses (NTP, 2015). • ***For Studies reporting SMRs or SIRs:*** No discussion of adjustments. Results are not adjusted for both age and sex (or stratified) if applicable.	
Not rated/applicable	• Do not select for this metric.	
Reviewer's comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*	
Metric 10. Covariate Characterization (measurement/information, confounding biases)		
High (score = 1)	• ***For all study types:*** Potential confounders (excluding co-exposures; e.g., age, sex, SES) were assessed using valid and reliable methodology where appropriate (e.g., validated questionnaires, biomarker).	
Medium (score = 2)	• ***For all study types:*** A less-established method was used to assess confounders (excluding co-exposures) and no method validation was conducted against well-established methods, but there was little to no evidence that that the method had poor validity and little to no evidence of confounding.	
Low (score = 3)	• ***For all study types:*** The confounder (excluding co-exposures) assessment method is an insensitive instrument or measure or a method of unknown validity.	
Unacceptable (score = 4)	• ***For all study types:*** Confounders were assessed using a method or instrument known to be invalid.	
Not rated/applicable	• ***For all study types:*** Covariates were not assessed.	
Reviewer's comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*	
Metric 11. Co-exposure Confounding (measurement/information, confounding biases)		
High (score = 1)	• Do not select for this metric.	
Medium (score = 2)	• ***For all study types:*** Any co-exposures to pollutants that are not the target exposure that would likely bias the results were not likely to be present. **OR** • Co-exposures to pollutants were appropriately measured or either directly or indirectly adjusted for.	

Confidence Level (Score)	Description	Selected Score
Low (score = 3)	• ***For cohort and cross-sectional studies:*** There is direct evidence that there was an unbalanced provision of additional co-exposures across the primary study groups, which were not appropriately adjusted for. • ***For case-control studies:*** There is direct evidence that there was an unbalanced provision of additional co-exposures across cases and controls, which were not appropriately adjusted for, and significant indication a biased exposure-outcome association.	
Unacceptable (score = 4)	• Do not select for this metric.	
Not rated/applicable	• Enter 'NA' and do not score this metric.	
Reviewer's comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*	
Domain 5. Analysis		
Metric 12. Study Design and Methods		
High (score = 1)	• Do not select for this metric.	
Medium (score = 2)	• ***For all study types:*** The study design chosen was appropriate for the research question (e.g., assess the association between exposure levels and common chronic diseases over time with cohort studies, assess the association between exposure and rare diseases with case-control studies, and assess the association between exposure levels and acute disease with a cross-sectional study design). **AND** • The study uses an appropriate statistical method to address the research question(s) (e.g., repeated measures analysis for longitudinal studies, logistic regression analysis for case-control studies, or mean, median for descriptive studies)	
Low (score = 3)	• Do not select for this metric.	
Unacceptable (score = 4)	• ***For all study types:*** The study design chosen was not appropriate for the research question. **OR** • Inappropriate statistical analyses were applied to assess the research questions.	
Not rated/applicable	• Do not select for this metric.	
Reviewer's comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*	
Metric 13. Statistical power (sensitivity)		
High (score = 1)	Do not select for this metric.	
Medium (score = 2)	• ***For cohort and cross-sectional studies:*** The number of participants are adequate to detect an effect in the exposed population and/or subgroups of the total population. **OR** • The paper reported statistical power is high enough (≥ 80%) to detect an effect in the exposure population and/or subgroups of the total population.	

Confidence Level (Score)	Description	Selected Score
	• ***For case-control studies:*** The number of cases and controls are adequate to detect an effect in the exposed population and/or subgroups of the total population. **OR** • The paper reported statistical power is high enough (≥ 80%) to detect an effect in the exposure population and/or subgroups of the total population.	
Low (score = 3)	• Do not select for this metric.	
Unacceptable (score = 4)	• ***For cohort and cross-sectional studies:*** The number of participants is inadequate to detect an effect in the exposed population and/or subgroups of the total population • ***For case-control studies:*** The number of cases and controls is inadequate to detect an effect in the exposed population and/or subgroups of the total population	
Not rated/applicable	• Do not select for this metric.	
Reviewer's comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*	
Metric 14. Reproducibility of analyses [adapted from Blettner et al. (2001)]		
High (score = 1)	• Do not select for this metric.	
Medium (score = 2)	• ***For all study types:*** The description of the analysis is sufficient to understand precisely what has been done and to be conceptually reproducible with access to the analytic data.	
Low (score = 3)	• ***For all study types:*** The description of the analysis is insufficient to understand what has been done and to be reproducible OR a description of analyses are not present (e.g., statistical tests and estimation procedures were not described, variables used in the analysis were not listed, transformations of continuous variables (e.g., logarithmic) were not explained, rules for categorization of continuous variables were not presented, exclusion of outliers was not elucidated and how missing values are dealt with was not mentioned).	
Unacceptable (score = 4)	• Do not select for this metric.	
Not rated/applicable	• Do not select for this metric.	
Reviewer's comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*	
Metric 15. Statistical Models (confounding bias)		
High (score = 1)	• Do not select for this metric.	
Medium (score = 2)	• ***For all study types:*** The model or method for calculating the risk estimates (e.g., odds ratios, SMRs, SIR) is transparent (i.e., it is stated how/why variables were included or excluded). • **AND** • Model assumptions were met.	
Low (score = 3)	• ***For all study types:*** The statistical model building process is not fully appropriate **OR** model assumptions were not met **OR** a description of analyses are not present [STROBE Checklist 12e (Von Elm et al., 2008)].	
Unacceptable	• Do not select for this metric.	

Confidence Level (Score)	Description	Selected Score
(score = 4)		
Not rated/applicable	• Enter 'NA' if the study did not use a statistical model.	
Reviewer's comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*	
Domain 6. Other (if applicable) Considerations for Biomarker Selection and Measurement (Lakind et al., 2014).		
Metric 16. Use of Biomarker of Exposure (detection/measurement/information biases)		
High (score = 1)	• Biomarker in a specified matrix has accurate and precise quantitative relationship with external exposure, internal dose, or target dose. **AND** • Biomarker is derived from exposure to one parent chemical.	
Medium (score = 2)	• Biomarker in a specified matrix has accurate and precise quantitative relationship with external exposure, internal dose, or target dose. **AND** • Biomarker is derived from multiple parent chemicals.	
Low (score = 3)	• Evidence exists for a relationship between biomarker in a specified matrix and external exposure, internal dose or target dose, but there has been no assessment of accuracy and precision or none was reported.	
Unacceptable (score = 4)	• Biomarker in a specified matrix is a poor surrogate (low accuracy, specificity, and precision) for exposure/dose.	
Not rated/applicable	• Enter 'NA' and do not score the metric if no biomarker of exposure was measured.	
Reviewer's comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*	
Metric 17. Effect biomarker (detection/measurement/information biases)		
High (score = 1)	• Effect biomarker measured is an indicator of a key event in an adverse outcome pathway (AOP).	
Medium (score = 2)	• Biomarkers of effect shown to have a relationship to health outcomes using well validated methods, but the mechanism of action is not understood.	
Low (score = 3)	• Biomarkers of effect shown to have a relationship to health outcomes, but the method is not well validated and mechanism of action is not understood.	
Unacceptable (score = 4)	• Biomarker has undetermined consequences (e.g., biomarker is not specific to a health outcome).	
Not rated/applicable	• Enter 'NA' and do not score the metric if no biomarker of effect was measured.	
Reviewer's comments		
Metric 18. Method sensitivity (detection/measurement/information biases)		
High (score = 1)	• Do not select for this metric.	
Medium (score = 2)	• Limits of detection are low enough to detect chemicals in a sufficient percentage of the samples to address the research question. Analytical methods measuring biomarker are adequately reported. The limit of detection (LOD) and limit of quantification (LOQ) (value or %) are reported.	
Low (score = 3)	• Frequency of detection too low to address the research hypothesis. **OR**	

Confidence Level (Score)	Description	Selected Score
	LOD/LOQ (value or %) are not stated	
Unacceptable (score = 4)	• Do not select for this metric.	
Not rated/applicable	• Enter 'NA' and do not score the metric.	
Reviewer's comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*	
Metric 19. Biomarker stability (detection/measurement/information biases)		
High (score = 1)	• Samples with a known storage history and documented stability data or those using real-time measurements.	
Medium (score = 2)	• Samples have known losses during storage, but the difference between low and high exposures can be qualitatively assessed.	
Low (score = 3)	• Samples with either unknown storage history and/or no stability data for target analytes and high likelihood of instability for the biomarker under consideration	
Unacceptable (score = 4)	• Do not select for this metric.	
Not rated/applicable	• Enter 'NA' and do not score the metric if no biomarkers were assessed.	
Reviewer's comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*	
Metric 20. Sample contamination (detection/measurement/information biases)		
High (score = 1)	• Samples are contamination-free from the time of collection to the time of measurement (e.g., by use of certified analyte free collection supplies and reference materials, and appropriate use of blanks both in the field and lab). **AND** • Documentation of the steps taken to provide the necessary assurance that the study data are reliable is included.	
Medium (score = 2)	• Samples are stated to be contamination-free from the time of collection to the time of measurement. **AND** • There is incomplete documentation of the steps taken to provide the necessary assurance that the study data are reliable.	
Low (score = 3)	• Samples are known to have contamination issues, but steps have been taken to address and correct contamination issues. **OR** • Samples are stated to be contamination-free from the time of collection to the time of measurement, but there is no use or documentation of the steps taken to provide the necessary assurance that the study data are reliable.	
Unacceptable (4)	• There are known contamination issues and no documentation that the issues were addressed.	
Not rated/applicable	• Enter 'NA' and do not score the metric if no samples were collected.	
Reviewer's comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*	

Confidence Level (Score)	Description	Selected Score
Metric 21. Method requirements (detection/measurement/information biases)		
High (score = 1)	• Instrumentation that provides unambiguous identification and quantitation of the biomarker at the required sensitivity [e.g., gas chromatography/high-resolution mass spectrometry (GC–HRMS); gas chromatography with tandem mass spectrometry (GC–MS/MS); liquid chromatography with tandem mass spectrometry (LC–MS/MS)].	
Medium (score = 2)	• Instrumentation that allows for identification of the biomarker with a high degree of confidence and the required sensitivity [e.g., gas chromatography mass spectrometry (GC–MS), gas chromatography with electron capture detector (GC–ECD)].	
Low (score = 3)	• Instrumentation that only allows for possible quantification of the biomarker, but the method has known interferants [e.g., gas chromatography with flame-ionization detection (GC–FID), spectroscopy].	
Unacceptable (score = 4)	• Do not select for this metric.	
Not rated/applicable	• Enter 'NA' and do not score the metric if biomarkers were not measured.	
Reviewer's comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*	
Metric 22. Matrix adjustment (detection/measurement/information biases)		
High (score = 1)	• If applicable for the biomarker under consideration, study provides results, either in the main publication or as a supplement, for both adjusted and unadjusted matrix concentrations (e.g., creatinine-adjusted or specific gravity-adjusted and non-adjusted urine concentrations) and reasons are given for adjustment approach.	
Medium (score = 2)	• If applicable for the biomarker under consideration, study only provides results using one method (matrix-adjusted or not).	
Low (score = 3)	• If applicable for the biomarker under consideration, no established method for matrix adjustment was conducted.	
Unacceptable (score = 4)	• Do not select for this metric.	
Not rated/applicable	• Enter 'NA' and do not score the metric if not applicable for the biomarker or no biomarker was assessed.	
Reviewer's comments	*[Document concerns, uncertainties, limitations, and deficiencies and any additional comments that may highlight study strengths or important elements such as relevance]*	

References

Blettner, MH, C. Razum, O. (2001). Critical reading of epidemiological papers. A guide. Eur J Public Health. 11(1): 97-101.

Cooper, GL, R. Agerstrand, M. Glenn, B. Kraft, A. Luke, A. Ratcliffe, J. (2016). Study sensitivity: Evaluating the ability to detect effects in systematic reviews of chemical exposures. Environ Int. 92-93: 605-610. http://dx.doi.org/10.1016/j.envint.2016.03.017.

Lakind, JSS, J. Goodman, M. Barr, D. B. Fuerst, P. Albertini, R. J. Arbuckle, T. Schoeters, G. Tan, Y. Teeguarden, J. Tornero-Velez, R. Weisel, C. P. (2014). A proposal for assessing study quality: Biomonitoring, Environmental Epidemiology, and Short-lived Chemicals (BEES-C) instrument. Environ Int. 73: 195-207. http://dx.doi.org/10.1016/j.envint.2014.07.011; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4310547/pdf/nihms-656623.pdf.

NTP. (2015). Handbook for conducting a literature-based health assessment using OHAT approach for systematic review and evidence integration. U.S. Dept. of Health and Human Services, National Toxicology Program. http://ntp.niehs.nih.gov/pubhealth/hat/noms/index-2.html.

Shamliyan, TK, R. L. Dickinson, S. (2010). A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for diseases [Review]. J Clin Epidemiol. 63(10): 1061-1070. http://dx.doi.org/10.1016/j.jclinepi.2010.04.014.

Von Elm, EA, D. G. Egger, M. Pocock, S. J. Gøtzsche, P. C. Vandenbroucke, J. P. (2008). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol. 61(4): 344-349. https://hero.epa.gov/heronet/index.cfm/reference/download/reference_id/4263036.