Reproducibility of the blood and urine exposome: A systematic literature review and meta-analysis

Jantje Goerdten, Li Yuan, Inge Huybrechts, Vanessa Neveu, Ute Nöthlings, Wolfgang Ahrens, Augustin Scalbert, Anna Floegel

DOI
10.1158/1055-9965.EPI-22-0090

Published in
Cancer Epidemiology, Biomarkers & Prevention

Document version
Accepted manuscript

Online publication date
23 June 2022

Corresponding author
Jantje Goerdten

Citation
Goerdten J, Yuan L, Huybrechts I, Neveu V, Nöthlings U, Ahrens W, et al. Reproducibility of the blood and urine exposome: A systematic literature review and meta-analysis. Cancer Epidemiol Biomarkers Prev. 2022;31(9):1683-92.

Reproducibility of the blood and urine exposome – a systematic literature review and meta-analysis

Running Title: Reproducibility of the blood and urine exposome

Jantje Goerdten[1], Li Yuan[1], Inge Huybrechts[2], Vanessa Neveu[2], Ute Nöthlings[3], Wolfgang Ahrens[1], Augustin Scalbert[2] & Anna Floegel[1,4]

[1]Leibniz-Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany

[2]International Agency for Research on Cancer (IARC), Lyon, France

[3]Unit of Nutritional Epidemiology, Department of Nutrition and Food Sciences, Rheinische Friedrich-Wilhelms-University Bonn, Bonn, Germany

[4]Section of Dietetics, Faculty of Agriculture and Food Sciences, Hochschule Neubrandenburg - University of Applied Sciences, Neubrandenburg, Germany

Conflict of interest

The authors declare no potential conflicts of interest

Corresponding Author: Jantje Goerdten

ORCID: 0000-0002-7814-1802

Email address: goerdten@leibniz-bips.de, sec-epi@leibniz-bips.de

Mailing address: Achterstraße 30, 28359 Bremen, Germany

Phone number: +49 (0)421 218-56844

Abstract

Endogenous and exogenous metabolite concentrations may be susceptible to variation over time. This variability can lead to misclassification of exposure levels and in turn to biased results. To assess the reproducibility of metabolites, the intra-class correlation coefficient (ICC) is computed. A literature search in three databases from 2000 until May 2021 was conducted to identify studies reporting ICCs for blood and urine metabolites. This review includes 192 studies, of which 31 studies are included in the meta-analyses. The ICCs of 359 single metabolites are reported and the ICCs of 10 metabolites were meta-analysed. The reproducibility of the single metabolites ranges from poor to excellent and is highly compound dependent. The reproducibility of BPA (bisphenol A), MEP (mono-ethyl phthalate), MnBP (mono-n-butyl phthalate), MEHP (mono-2-ethylhexyl phthalate), MEHHP (mono(2-ethyl-5-hydroxyhexyl) phthalate), MBzP (mono-benzyl phthalate), MEOHP (mono-(2-ethyl-5-oxohexyl) phthalate), methylparaben and propylparaben, is poor to moderate (ICC median: 0.32; range: 0.15-0.49) and for 25-Hydroxyvitamin D (25(OH)D) excellent (ICC: 0.95; 95% CI: 0.90, 0.99). Pharmacokinetics, mainly the half-life of elimination and exposure patterns, can explain reproducibility. This review describes the reproducibility of the blood and urine exposome, provides a vast dataset of ICC estimates, and hence constitutes a valuable resource for future reproducibility, clinical and epidemiological studies.

Keywords: reproducibility, reliability, metabolites, intra class correlation coefficient, variability, exposome

1 Introduction

The exposome is the totality of all exposures accumulating during a person's lifetime, including not only exogenous exposures but also exposures from endogenous processes (1). There is now a wealth of biomarkers available measuring the exposome, thanks to the progress of mass spectrometry and the development of metabolomics (2, 3). In metabolomics small molecules (molecular mass of 50-1500 Da) involved in the metabolism are measured, which are called metabolites. These biomarkers are intended to be used in epidemiological studies for exposure assessment (4). Though there is a high demand for biomarkers of exposure, the reproducibility of these biomarkers is widely unknown. Due to the scattered information regarding the reproducibility of metabolite concentrations.

There are different sources of variability, specifically: nature of biospecimen, time of sampling, mode of collection and storage, within-subject variation over time and laboratory error (5). Epidemiological studies apply predominantly single biomarker measurements to reflect long-term exposure (6). However, metabolite concentrations may be susceptible to variation over time. This variability can lead to biased results, namely, non-differential misclassification bias, which moves the risk estimate towards the null, i.e., no effect is found even though one is present (5). Hence, the variability of a biomarker must be evaluated, and its sources identified when assessing the appropriateness of a biomarker to be used in epidemiological studies (7).

Reproducibility, sometimes called reliability, is the term to describe the variation between two measurements made on the same subject under varying conditions, e.g., repeated measurements on the same sample, or measurements of samples collected at two different time points (8, 9). A biomarker needs to be reasonably stable over time, meaning, to show high reproducibility in an individual. This translates to a smaller within-subject variance over time compared to the between-subject variance, which should explain most of the variation seen in the measurement (10, 11). The intra-class correlation coefficient (ICC) is used to assess reproducibility and

equals the correlation between any two or more measurements made on the same subject (8, 12). It is generally computed by dividing the between-subject variation by the total variation (sum of the within-subject variation and between-subject variation) (12). The ICC can take any value between 0 and 1, and is commonly interpreted as < 0.4 poor, ≥ 0.4 moderate, ≥ 0.7 good and ≥ 0.85 excellent reproducibility (13). The ICC can be an important parameter to increase understanding of the variability of a given metabolite in a specific type of biospecimen.

Only a few reviews have been conducted, collating studies on the reproducibility of biomarkers. One systematic review and meta-analyses included 368 studies to assess the reproducibility of hormone concentrations in blood and other biospecimens (i.e., urine, saliva, faeces etc.) (14). The authors found moderate (ICC 0.68) reproducibility of hormone levels in human studies. Another systematic review collected evidence on the reproducibility of whole-grain and cereal fibre intake biomarkers (15). The authors concluded that the medium- to long-term reproducibility of these biomarkers was poor and a substantial limitation for their clinical use. Furthermore, two studies assessed the reproducibility of urinary biomarkers of exposure to non-persistent chemicals, such as phthalates (16, 17). The authors found most biomarkers to have low reproducibility (ICCs <0.4) and only 6% of biomarkers showed high reproducibility (ICCs >0.75) (16). Also, the authors of the other study found low reproducibility of phthalate biomarkers (ICC ranging from 0.1 to 0.6) (17). Another systematic literature review summarised the reproducibility of triclosan, where the ICC ranges from 0.3 to 0.9 in the included studies (18). Additionally, *Exposome-Explorer* is a database aiming to collect and summarise comprehensive data on all known biomarkers of exposure, including information on reproducibility (19, 20).

To our knowledge, no further reviews summarising the reproducibility of metabolite concentrations were conducted in the past. There is a gap in the literature to extensively summarise the existing evidence on the reproducibility of urine and blood metabolite

concentrations constituting the exposome. The increasing number of available/ discovered biomarkers and interest in using these in clinical settings and research shows the great need to assess the reproducibility of these biomarkers and summarise possible biological variability over time. Hence, this review aims to summarise the present literature on the reproducibility of urinary and blood metabolite concentrations, present meta-analyses of ICC estimates, and provide useful guidance for future studies.

2 Materials and Methods

2.1 Search Strategy

We searched PubMed, ISI Web of Science and Scopus for relevant articles, from 1st January 2000 to May 5th, 2021. The authors used the following search terms and variations of those: "reproducibility of results", "biological variation", "variability", "reliability", "stability", "reproducibility", "within person variation", "between person variation", "intra class correlation coefficient", "biomarker", "metabolomics", "metabolome", "blood", "serum", "plasma", "urine" and "metabolites". Whenever possible, these terms were mapped to Medical Subject Headings (MeSH) (Supplementary Materials and Methods 1 contains the full search strategy).

2.2 Study selection for qualitative synthesis

First, two independent reviewers (JG and LY) screened the titles and abstracts of the non-duplicated references retrieved from the databases. Second, full-text articles of the selected references were screened for eligibility by one reviewer (JG). The following in- and exclusion criteria were used during the screening process: (1) studies with assessing reproducibility as part of the main objectives; (2) at least two measurements taken from a subject; (3) more than

one time point assessed; (4) metabolites assessed from urine or blood; (5) the study was conducted in humans; (6) the study had to be published in a peer-reviewed journal and English language. Conference abstracts, short communications, editorials, or comments were excluded, due to the limited information available in the text. Any disagreements during the screening process were resolved by consensus between the two reviewers (JG and LY), or if necessary, by a third independent reviewer (AF) if the disagreement could not be resolved. This review was conducted following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (21). The review protocol was not prior published.

2.3 Data extraction

Data were extracted from each eligible full text by one author (JG). Information collected includes data source, study population, sample size, subject characteristics, study time, time points of collection, number of samples taken at each time point, state at collection (fasting, phase of menstruation cycle, etc.), samples from urine or blood, metabolites collected, metabolite platform, ICCs, adjustments made to the ICC, ICC formula, statistical technique used for ICC calculation, and ICC classification scheme. Information was collected for each metabolite alone and not a combination/ summary of the study, i.e., studies looking at the same metabolite under different conditions will have several rows to summarise the information of the metabolite in each condition. Except if a study assessed the reproducibility of more than 70 metabolites, then only a summary of the study was reported, e.g., average or range of ICCs for groups of metabolites, classes, or all metabolites. ICCs are reported as adjusted ICCs whenever possible.

2.4 Quality assessment and risk of bias

The quality of the included references in this review was assessed by two independent reviewers (JG and LY). As no adequate quality assessment tool for our purposes was available a combination of tools from the Biomonitoring, Environmental Epidemiology, and Short-Lived Chemicals (BEES-C) (22) instrument and the quality assessment tool designed to assess biomarker-based cross-sectional studies (BIOCROSS) (23) was created and defined for reproducibility studies of metabolites in humans. Even though no cross-sectional studies are included in this review, the BIOCROSS tool provides unique issues to consider for studies including biomarkers. Likewise, the BEES-C tool is developed to assess the quality of epidemiological studies involving biomonitoring of chemicals with short physiological half-lives, however, this tool adds additional aspects unique to biomarker studies (Supplementary Table S1). The total score is 24, the overall study quality is based on the awarded score and is regarded as ≤7 poor, ≥8 - ≤16 fair, ≥17 good study quality.

2.5 Study selection for quantitative synthesis

Based on the previous selected studies for the quantitative synthesis, the authors reduced the number of metabolites available for analysis by selecting only metabolites occurring ≥ 10 times in the extracted data, i.e., having ≥ 10 ICC estimates available. Furthermore, for the analysis the exact number of specimens taken, and the number of time points is crucial information, when not available, the study is excluded from the analysis. If after following these exclusion criteria the number of ICC estimates is reduced to under 10, then the metabolite is excluded as well. For urine metabolites the ICC was only included in the analysis when the concentrations were adjusted for dilution (e.g., specific gravity, creatinine, or osmolality).

2.6 Data analysis

All analyses were performed in R Studio (Version 4.0.3) (24) and the packages *metafor* (25), *foreach* (26) and *ggplot2* (27) were used. Normally when meta-analysing correlation coefficients, the coefficients are transformed to the Fisher's z scale. This is done because the variance depends strongly on the correlation (28). However, no straightforward method exists to convert z-transformed ICC values back to an interpretable ICC after the meta-analysis. Hence, we applied a previously tested method, where we did not standardise the ICC estimates and used the raw estimates (29). For this method, a normal distribution of the ICC estimates is assumed, and the sample variance could be approximated with the following equation (Eqn. 1).

$$Var\ ICC = \frac{2 \times (1 - ICC)^2 \times (1 + (n-1) \times ICC)^2}{k \times n \times (n-1)}\ (Eqn.1)$$

Where *ICC* is the raw ICC estimate extracted from the full text, *k* is the number of repeated measures per subject and *n* is the number of individuals in the study.

Random effects models were carried out to obtain summary ICC estimates for each metabolite included in the quantitative synthesis. The model is modelled with a random effect for authors and a second-level random effect, as higher clustering is present since studies reported several ICCs for one metabolite. Furthermore, to assess heterogeneity, i.e., if the effect sizes are consistent across studies, we estimated the I^2 from the computed random effect models (30).

Publication bias can be present, as studies that report significant effect sizes are more likely to be published than studies that have smaller effect sizes or no significant findings at all (28). To evaluate the presence of publication bias we computed Kendall's Tau and visually assessed the computed funnel plots.

2.6.1 Visualisation

The metabolites with available reproducibility data were plotted as a similarity network based on chemical structures. The visualization was accomplished using Cytoscape, a software dedicated to the visualization and analysis of complex biological networks (31).

The chemical structures were retrieved from PubChem (32) using the *get_cid()* and *pc_prop()* functions from the R *webchem* package (33). Compound names were used to get the PubChem IDs. These IDs were used to retrieve the chemical data (isomeric simplified molecular-input line-entry system (SMILES)).

The network calculation and figure generation in Cytoscape were automated with the *RCy3* package (34, 35). The similarity network was computed with a 0.8 Tanimoto coefficient using the *chemViz2* Cytoscape app (36).

2.7 Data availability

The data generated in this study for the meta-analyses are available upon request from the corresponding author. The data generated in this study, apart from the data used for the meta-analyses, are available within the article and its supplementary data files.

3 Results

The literature search resulted in 13,536 records. After duplicate removal 10,185 records were screened, of which 9,944 were excluded, resulting in 241 records eligible for full-text screening. 15 full-text articles were added to the full-text screening from the database Exposome Explorer (19, 20). The main reason for exclusion was that no metabolites (e.g., cytokines) were assessed in the study, followed by reproducibility not being one of the main objectives of the study. In the end, 192 studies were included in the qualitative synthesis and of these 31 studies were included in the meta-analyses (Figure 1(21)). The study quality is overall high, indicating a low risk of bias. The mean score for all included studies is 16.9 (SD= 1.9) and the median score is 17 (range= 10-21). Most studies (N= 119) are in the high quartile with a score of ≥17 and 73 studies are in the medium quartile with a score of ≥8 and <17 (Supplementary Materials and

Methods 2). For the domain “Reliability and reproducibility specific considerations” with max. 3 points, the average for all studies is 2.74. Further, for the domain “Biomarker measurement” with max. 2 points, the average is 1.63, for the domain “Specimen characteristics and essay methods” with max. 3 points, the average is 1.32, and for the domain “Laboratory measurements” with max. 3 points, the average is 1.70. The study quality for the sub-domains “Reliability” and “Biomarker” is overall good, while for “Specimen” and “Laboratory” the quality can be considered fair.

The sample size of the included studies ranges from five to 3,455 individuals and the study time ranges from one day to up to 15 years. Five studies did not sufficiently report the study time. Almost all studies have a study time below 10 years, only three studies have a study time above 10 years and 106 studies have a study time ≤ 6 months. The total number of samples per subject ranges from two to up to 65, while three studies did not report the total number of samples per participant. The time points when specimens were collected ranges from two to 46 time points of the collection over the study period, while three studies did not report the time points. The included studies based their analysis on diverse study populations such as children, pregnant women, pre- and postmenopausal women, elderly, and patients with chronic diseases. Specimens were collected under 13 different sampling conditions, for example, fasting and non-fasting, or at some (luteal or follicular) phase during the menstrual cycle. The full table of extracted data for all included studies is presented in Supplementary Materials and Methods 2. Supplementary Materials and Methods 2 is intended as an interactive dataset, for researchers to search and select specific metabolites or chemical classes and to explore the corresponding reproducibility studies.

3.1 Blood and urine metabolites

In total 359 single metabolites (i.e., no classes/ groups/ ∑ or other summed metabolites) are analysed in the included studies, and 98 classes and summed metabolites were additionally recorded. A total of 14 studies either included ≥ 70 metabolites or only reported a median ICC for groups of metabolites or all analysed metabolites (11, 37-50). Of the 359 single metabolites, 97 were only analysed in blood and 216 only in urine, and 46 were analysed in blood and urine. *Benzene and substituted derivatives* (n= 70), *fatty acyls* (n= 38) and *carboxylic acids and derivatives* (n= 25) are the top three of the most common metabolite classes (defined from The Human Metabolome Database (HMDB) (51-54) and *Exposome Explorer*). In sum 67 different metabolite classes are included in this review. Several metabolites (n= 58/ 359) could not be classified from HMDB or the *Exposome Explorer* database. The most widely studied metabolites are bisphenol A (BPA, n= 51 in a total of 33 studies), mono(2-ethyl-5-hydroxyhexyl) phthalate (MEHHP, n= 49 in a total of 30 studies) and mono-ethyl phthalate (MEP, n= 48 in a total of 33 studies). 183 metabolites only occur once in the included studies, i.e., only one ICC estimate is reported.

In Supplementary Table S2, all 359 single metabolites are summarised, providing information on the chemical class, use, biosample and a list of all available ICC estimates for the specific metabolite. The classified uses of the metabolites range from oxidative stress markers, dietary or metal compounds, to environmental toxicants, pesticides/insecticides/herbicides, plasticizers, or antibiotics. Bringing together a part of the blood and urine exposome.

The metabolites, that could be identified in PubChem and where the chemical class membership is available (N=352), are visualised in two different figures (Fig. 2-3). The lines connecting compounds, called edges, represent the similarity between metabolites in the two figures. The sizes of the nodes are proportional to the number of ICC estimates available for the specific compound.

In Figure 2 the metabolites are grouped according to chemical class and the colour of the nodes indicates category membership (diet, endogenous, pollutants and drugs) and their subcategories. The compounds with the largest number of ICC estimates are pollutants, more specifically the subcategory phthalates. Also widely studied is the group of dietary compounds, however, the number of ICC estimates per compound is lower.

In Figure 3 the metabolites are grouped by chemical class and the colours of the node indicate the mean ICC classification. Where a darker shade indicates a higher average ICC estimate for the compound and a lighter shade a lower average ICC estimate. Only a few nodes, meaning compounds, have a dark colour. Especially the larger nodes show the lightest shades, and hence, only low reproducibility. Compounds belonging to the class of fatty acyl have overall the highest average ICC estimates, followed by steroids and derivatives.

In Supplementary Figure S1 the metabolites are grouped according to the chemical class, while here the colours of the nodes show the proportion of ICC estimates derived from urine or blood compounds from the total number of ICC estimates available for the compound.

3.1 ICC calculation and interpretation

The ICC is commonly calculated by the following equation (Eqn. 2), where $\sigma^2_{between}$ (σ^2_b) and σ^2_{within} (σ^2_w) are between-subject variance and within-subject variance, respectively.

$$ICC = \frac{\sigma_b^2}{(\sigma_b^2 + \sigma_w^2)} \quad (Eqn. 2)$$

In this review 114 studies apply Eqn. 2 to calculate ICCs, seven studies employ an alternative formula, while 71 studies do not report the applied ICC formula. A study (55) only reported the values for between- and within-subject variance, hence the ICC estimates were manually calculated by applying Eqn. 2. The required variance estimates for the calculation of the ICC

need to be computed by a statistical model. In total 33 different methods were used, here most studies applied a linear mixed model (n= 74/ 192), followed by an analysis of variance model (ANOVA) (n= 44/ 192), and random effects model (n= 26/ 192). Again, some studies (n= 38/ 192) do not report the employed statistical method. From the extracted data three types of adjustments could be identified: urine concentration (creatinine, specific gravity, or osmolality), time of sampling (fasting/non-fasting, season, etc.) and individual characteristics (age, sex, body-mass-index, etc.). The adjustments made to the urine concentration, are made before the calculation of the ICC estimate and should be systematically applied. However, 34 studies did not report an adjustment made to the urine concentration. Whereas the two other types are made during the calculation of the ICC estimate.

3.3 Meta-analyses results

In total 10 metabolites from 31 studies were included in the meta-analyses. All metabolites, except 25-Hydroxyvitamin D (25(OH)D), which belongs to the *prenol lipids* class and is measured in blood, belong to the *benzene and substituted derivatives* class and are measured in urine. Ten or more creatinine adjusted ICC estimates for BPA, MEP, mono-n-butyl phthalate (MnBP), mono-2-ethylhexyl phthalate (MEHP), MEHHP, mono-benzyl phthalate (MBzP), mono-(2-ethyl-5-oxohexyl) phthalate (MEOHP), methylparaben and propylparaben are available from the included studies. Only unadjusted ICC estimates are reported for 25(OH)D in the included studies. The reproducibility measurements are mainly based on samples from women and the longest study time is up to 8 years in the included studies. An overview of the studies included in the meta-analyses is presented in Supplementary Table S3-21. The results of the 9 metabolites where creatinine adjusted ICCs were available, and of the one metabolite of vitamin D are presented in Table 1, the corresponding forest plots are in Supplementary Fig. S2-38. Visual assessment of the computed funnel plots (Supplementary Fig. S3-39) and

Kendall's Tau values (Table 1) are indicative of the presence of publication bias for BPA, MEHHP, MEOHP (adjusted), and 25(OH)D. Furthermore, the results for MEP, MnBP, MEHHP, MEOHP, and propylparaben are highly inconsistent ($I^2 \geq 75\%$, Table 1).

The reproducibility of BPA, MnBP, MEHP, MEHHP, MBzP and MEOHP adjusted for creatinine can be classified as poor (ICC < 0.4), for MEP, methylparaben and propylparaben adjusted for creatinine as moderate (ICC > 0.4), and for 25(OH)D as excellent (ICC > 0.9).

Results for the unadjusted concentrations of BPA, MnBP, MEHP, MEHHP, MBzP, BP-3, MEOHP, MEP, methylparaben and propylparaben can be found in Supplementary Table S4-23 and Fig. S4-43. The meta-analysed ICC estimates are generally higher compared to the adjusted analyses.

4 Discussion

In this review, we compiled all suitable studies investigating the reproducibility of the blood and urine exposome. This results in the formation of a dataset containing the ICC estimates for 359 single metabolites and further 98 classes and summed metabolites. Additionally, an overview of the study and metabolite specific information is provided. The meta-analyses of the ICC estimates of 10 metabolites, showed low to moderate reproducibility for the 9 non-persistent chemicals and high reproducibility for a persistent metabolite of vitamin D.

4.1 Sources of Variability of ICC estimates

The observed variability between ICC estimates for the blood and urine exposome can have several sources: the nature of the biospecimen, the time of collection, the time intervals between the collection, and the population from which biospecimens are collected (10). For example, two studies compare the reproducibility of metal compound and essential element

measurements in spot, FMV and 24h urine samples collected in adult men (56,57). The differences in reproducibility estimates between biospecimens were highly dependent on the nature of the compound. Differences between FMV, spot and 24h urine samples, or blood can be explained by varying exposures over time and/or half-lives of elimination, i.e., the rate at which the exposure is cleared from the body. Thus, the observed reproducibility of a specific metabolite can depend on the timing of collection. Fluctuating variability patterns in populations over time can further explain differences between ICC estimates. However, the information on which biospecimen, or sampling time, results in the most reliable metabolite concentration is highly compound dependent. Hence, it is nearly impossible to give overall recommendations for (classes of) metabolites. Some metabolite concentrations present low variability over time, are reproducible in different populations or show limited differences in variability when measured in blood or urine. On the other hand, some metabolites might show distinct variabilities in different populations, due to varying exposure patterns, i.e., the source of exposure is not always present. Another problem could be that the metabolite is rapidly excreted from the body, resulting in large variability along time, unless the exposure source frequently reappears. In summary, the reproducibility of metabolite concentrations depends on two main factors: pharmacokinetics, mainly half-life of elimination, and frequency of exposure. Another aspect is the adjustment for urine dilution, some studies did not report any adjustments made to the urine samples during laboratory analysis. It is not clear if these studies did not adjust for urine dilution or plainly did not report it, all the same, ICC estimates based on adjusted urine samples tend to be overall lower, as seen in the meta-analyses results. Hence this needs to be considered when comparing ICC estimates.

It is not possible to give overall statements regarding the reproducibility of all blood or urine metabolite concentrations. Thus, to meta-analyse ICC estimates can provide a general idea of the reproducibility of a specific metabolite. This can offer useful guidance when planning a study with metabolite-based exposure assessment. However, whenever possible ICC estimates

derived from studies in a similar population should be assessed for the planning of a study measuring the specific metabolite (for this Supplementary Materials and Methods 2 can be used).

4.2 ICC formula and calculation

The studies included in this review applied multiple different statistical approaches to compute ICC estimates. In a recent paper, the authors (12) compare three statistical methods (restricted maximum likelihood from a linear mixed model, ANOVA and a variance estimate method) to compute the ICC from synthetic biomonitoring data. They find no major differences between the three analytical techniques, and the results stay the same even under changing conditions, i.e., missing values, suboptimal distributions, unbalanced data sets and unusual variance estimates. Hence there should be no major differences in the ICC estimates computed by different statistical approaches in the literature. However, there are studies in this review that applied uncommon statistical techniques or formulas to calculate ICC estimates. A novel analytical technique to calculate the ICC (58) is employed by one of the included studies (59). One study (60), calculated the so-called ICC(2,1) (61) and another study (49) included the technical variance in the ICC formula. Here it is not clear if these estimates are comparable with ICC estimates derived from standard statistical techniques and the common formula (Eqn. 2). We must point out that these approaches hamper the comparability with other studies. Similarly, one study (62) refrained from adjusting the ICC for fixed factors such as age and sex, due to the fear of decreasing comparability. It is possible to adjust for potential covariate influences and by this remove within-subject variation that arises from individual characteristics at the time of collection, however unrelated to exposure. Applying a novel statistical technique to calculate ICC estimates or to adjust for individual characteristics can enhance the understanding of the variability of metabolite concentrations; however,

comparability needs to be considered. We recommend when presenting adjusted ICCs to display unadjusted values as well. Furthermore, when applying a novel analytical technique, we advise including, whenever possible, ICC estimates obtained from a standard technique, e.g., ANOVA. This way comparability can be increased and differences in methods can be better understood.

4.3 Meta-analyses of ICC estimates

The 25(OH)D metabolite measured in blood is the only metabolite that can be classified as highly reproducible. MEP, BP-3, MBzP, MiBP, methylparaben and propylparaben, measured in urine, can be classified as moderately reproducible. BPA, MnBP, MEHP, MEHHP and MEOHP, measured in urine, are all poorly reproducible. For BPA, MEP, MnBP, MEHHP, MEOHP, propylparaben, methylparaben and 25(OH)D are either large heterogeneity, publication bias or both present. Hence, the results for these metabolites need to be interpreted with caution. One review showed similar results for non-persistent chemicals (including BPA, MEP, MnBP, MEHP, BP-3, MEHHP, MBzP, MEOHP; MiBP, methylparaben and propylparaben), most ICCs fall under the categories indicating only poor to moderate reproducibility (16). Furthermore, the authors also state the great inconsistency in the results of the included studies and attribute this to different parameters. The low reproducibility of these chemicals might be due to short half-lives and/or varying exposure patterns. As the authors already state in their conclusion, these results show the necessity for multiple samples per subject when measuring these metabolites in a study (16).

4.4 Repeated measurements

For metabolites with higher reproducibility, single measurements can be acceptable to depict long-term exposure, that is the “usual level” of exposure (6). As these metabolite concentrations

show higher between-individual variability than within-individual variability. Making single measurements of these metabolite concentrations adequate at the relative ordering of individual exposure levels (63). On the other hand, ICC estimates below 0.6 have been found to bias results (64). Many metabolite concentrations in this review show low ICC estimates, indicating poor to moderate reproducibility (Fig. 2 and Supplementary Materials and Methods 2). When using these metabolites as biomarkers of exposure the number of measurements per subject needs to be increased, to reduce the impact of low reproducibility. Yet, it is not always possible to increase the specimen collection per subject, due to cost restrictions or strain on the subjects. Here are three possible, however not exhaustive, approaches to handle this problem: In one study the authors propose a statistical method to estimate lifetime exposures from spot biomarkers using ICC estimates (6). The authors present a way to improve spot measurement-based risk estimates, by using ICC estimates from the literature, or if feasible collecting repeated measurements for a small subsample and calculating ICCs based on the collected data. Alternatively, another study presents an approach, when it is possible to collect several repeated measurements per subject, but cost restrictions are in place (64). The authors propose within-subject pooling of biospecimen samples before laboratory analysis. This method can reduce laboratory costs and the authors show, that increasing the measurements per subject and pooling them, is efficient in decreasing bias and increasing statistical power without affecting assay costs. Correspondingly, power calculations are often based on time-invariant exposures, however, this is mostly not the case in observational studies (65). Authors of a study developed a power calculation method where exposure variability and the costs of repeated measurements are taken into account (65). This way the number of participants and the number of measurements, while accounting for the total cost of the study, can be explored to optimize the power of a study.

4.5 Strengths and limitations

This review is an extensive summary of the existing literature presenting ICC estimates for the blood and urine exposome. We applied broad inclusion criteria, allowing a comprehensive collection of available ICC estimates for a great number of metabolite concentrations. However, there are still more ICC estimates available in the literature, as some studies additionally report computed ICCs. These studies were mainly excluded as reproducibility was not part of the main objectives of the study. The presented method to meta-analyse ICC estimates is not optimal, and some of the results show great heterogeneity and publication bias is present. Further, these analyses are only carried out for ICC estimates from studies presenting the required information, where several studies could not be included in the analysis, which could have potentially reduced heterogeneity and publication bias. This is the first attempt to meta-analyse ICC estimates from such a variety of metabolite concentrations. Further work into the best methodology to meta-analyse ICC estimates is needed.

5 Conclusion

This review collected the ICC estimates of 359 single exogenous and endogenous metabolite concentrations, and of additionally almost 100 classes of (or summed) metabolites. Making this review one of the first comprehensive reviews summarising the available information about the reproducibility of the blood and urine exposome. The results from the meta-analyses give a first indication of the general reproducibility of 9 non-persistent chemicals and one persistent metabolite of vitamin D. Moreover, further aspects of variability are discussed, and recommendations to handle low reproducibility are given. The vast dataset of information on the reproducibility of the exposome can be used by researchers to help interpret findings and to plan biospecimen collection. This makes this review a useful source for future reproducibility studies and epidemiological studies planning to use metabolite-based (exposure) assessment.

Acknowledgements

Funding

This study was funded by the German Research Foundation (FL 884/3-1) the recipient is A. Floegel, and the Agence Nationale de la Recherche the recipient is A. Scalbert.

Author Contributions

Jantje Goerdten and Anna Floegel were involved in the conceptualization of this manuscript. Jantje Goerdten performed the literature search, screening process, statistical analysis and wrote the manuscript. Anna Floegel was the main supervisor and reviewed and edited the manuscript. Augustin Scalbert provided major theoretical input and reviewed and edited the manuscript. Li Yuan validated the screening process, reviewed, and edited the manuscript. Vanessa Neveu visualised the data, reviewed and edited the manuscript. Ute Nöthlings, Inge Huybrechts, Wolfgang Ahrens reviewed and edited the manuscript.

References

1. Rappaport SM. Implications of the exposome for exposure science. Journal of Exposure Science & Environmental Epidemiology **2011**;21(1):5-9 doi 10.1038/jes.2010.50.
2. Sengupta A, Narad P. Metabolomics. In: Arivaradarajan P, Misra G, editors. Omics Approaches, Technologies And Applications: Integrative Approaches For Understanding OMICS Data. Singapore: Springer Singapore; 2018. p 75-97.
3. Zhang A, Sun H, Wang X. Emerging role and recent applications of metabolomics biomarkers in obesity disease research. RSC Advances **2017**;7(25):14966-73 doi 10.1039/C6RA28715H.
4. Dragsted LO, Gao Q, Praticò G, Manach C, Wishart DS, Scalbert A*, et al.* Dietary and health biomarkers-time for an update. Genes Nutr **2017**;12:24 doi 10.1186/s12263-017-0578-y.
5. White E. Measurement error in biomarkers: sources, assessment, and impact on studies. IARC Sci Publ **2011**(163):143-61.
6. Pleil JD, Sobus JR. Estimating Lifetime Risk from Spot Biomarker Data and Intraclass Correlation Coefficients (ICC). Journal of Toxicology and Environmental Health-Part a-Current Issues **2013**;76(12):747-66 doi 10.1080/15287394.2013.821394.
7. Kuhnle GGC. Stable isotope ratios - nutritional biomarkers of long-term intake? American Journal of Clinical Nutrition **2019**;110(6):1265-7 doi 10.1093/ajcn/nqz239.
8. Bartlett JW, Frost C. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. Ultrasound in Obstetrics & Gynecology **2008**;31(4):466-75 doi 10.1002/uog.5256.
9. Plesser HE. Reproducibility vs. Replicability: A Brief History of a Confused Terminology. Frontiers in Neuroinformatics **2018**;11(76) doi 10.3389/fninf.2017.00076.
10. Sampson JN, Boca SM, Shu XO, Stolzenberg-Solomon RZ, Matthews CE, Hsing AW*, et al.* Metabolomics in epidemiology: sources of variability in metabolite measurements and implications. Cancer Epidemiol Biomarkers Prev **2013**;22(4):631-40 doi 10.1158/1055-9965.Epi-12-1109.
11. Floegel A, Drogan D, Wang-Sattler R, Prehn C, Illig T, Adamski J*, et al.* Reliability of serum metabolite concentrations over a 4-month period using a targeted metabolomic approach. PLoS One **2011**;6(6):e21103 doi 10.1371/journal.pone.0021103.
12. Pleil JD, Wallace MAG, Stiegel MA, Funk WE. Human biomarker interpretation: the importance of intra-class correlation coefficients (ICC) and their calculations based on mixed models, ANOVA, and variance estimates. J Toxicol Environ Health B Crit Rev **2018**;21(3):161-80 doi 10.1080/10937404.2018.1490128.
13. Rosner B. Fundamentals of biostatistics. Nelson Education; 2015. p. 609-613.
14. Fanson KV, Biro PA. Meta-analytic insights into factors influencing the repeatability of hormone levels in agricultural, ecological, and medical fields. Am J Physiol Regul Integr Comp Physiol **2019**;316(2):R101-r9 doi 10.1152/ajpregu.00006.2018.
15. Jawhara M, Sørensen SB, Heitmann BL, Andersen V. Biomarkers of Whole-Grain and Cereal-Fiber Intake in Human Studies: A Systematic Review of the Available Evidence and Perspectives. Nutrients **2019**;11(12) doi 10.3390/nu11122994.
16. LaKind JS, Idri F, Naiman DQ, Verner MA. Biomonitoring and Nonpersistent Chemicals-Understanding and Addressing Variability and Exposure Misclassification. Curr Environ Health Rep **2019**;6(1):16-21 doi 10.1007/s40572-019-0227-2.
17. Johns LE, Cooper GS, Galizia A, Meeker JD. Exposure assessment issues in epidemiology studies of phthalates. Environ Int **2015**;85:27-39 doi 10.1016/j.envint.2015.08.005.
18. Goodman M, Naiman DQ, LaKind JS. Systematic review of the literature on triclosan and health outcomes in humans. Critical reviews in toxicology **2018**;48(1):1-51.
19. Neveu V, Moussy A, Rouaix H, Wedekind R, Pon A, Knox C*, et al.* Exposome-Explorer: a manually-curated database on biomarkers of exposure to dietary and environmental factors. Nucleic Acids Res **2017**;45(D1):D979-d84 doi 10.1093/nar/gkw980.

20. Neveu V, Nicolas G, Salek RM, Wishart DS, Scalbert A. Exposome-Explorer 2.0: an update incorporating candidate dietary biomarkers and dietary associations with cancer risk. Nucleic Acids Research **2020**;48(D1):D908-D12 doi 10.1093/nar/gkz1009.
21. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD*, et al.* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ **2021**;372:n71 doi 10.1136/bmj.n71.
22. LaKind JS, Sobus JR, Goodman M, Barr DB, Fürst P, Albertini RJ*, et al.* A proposal for assessing study quality: Biomonitoring, Environmental Epidemiology, and Short-lived Chemicals (BEES-C) instrument. Environment International **2014**;73:195-207 doi 10.1016/j.envint.2014.07.011.
23. Wirsching J, Graßmann S, Eichelmann F, Harms LM, Schenk M, Barth E*, et al.* Development and reliability assessment of a new quality appraisal tool for cross-sectional studies using biomarker data (BIOCROSS). BMC Med Res Methodol **2018**;18(1):122 doi 10.1186/s12874-018-0583-x.
24. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Version 4.0.3 [software]; **2018** [cited 2021 Oct 12]. Available from: https://www.R-project.org/.
25. Viechtbauer W. Conducting meta-analyses in R with the metafor package. Journal of statistical software **2010**;36(3):1-48.
26. Microsoft, Weston S. foreach: Provides Foreach Looping Construct. R package version 1.5.1 [software]. **2020** [cited 2021 Oct 12]. Available from: https://CRAN.R-project.org/package=foreach.
27. Wickham H. GGPLOT2: Elegant Graphics for Data Analysis 2016 Springer-Verlag, New York. R package version 3.3.5 [software]. **2016** [cited 2021 Oct 12].
28. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. Introduction to meta-analysis. John Wiley & Sons; 2011. p. 39-41.
29. Noble S, Scheinost D, Constable RT. A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. NeuroImage **2019**;203 doi 10.1016/j.neuroimage.2019.116157.
30. Viechtbauer W [Internet]. I^2 for Multilevel and Multivariate Models [cited 2021 Jun 15]. Available from: https://www.metafor-project.org/doku.php/tips:i2_multilevel_multivariate.
31. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D*, et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research **2003**;13(11):2498-504 doi 10.1101/gr.1239303.
32. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S*, et al.* PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Research **2021**;49(D1):D1388-D95 doi 10.1093/nar/gkaa971.
33. Szöcs E, Stirling T, Scott ER, Scharmüller A, Schäfer RB. Webchem: an r package to retrieve chemical information from the Web. Journal of Statistical Software **2020**;93(1):1-17.
34. Gustavsen JA, Pai S, Isserlin R, Demchak B, Pico AR. RCy3: Network biology using Cytoscape from within R. F1000Research **2019**;8.
35. Otasek D, Morris JH, Bouças J, Pico AR, Demchak B. Cytoscape Automation: empowering workflow-based network analysis. Genome biology **2019**;20(1):185 doi 10.1186/s13059-019-1758-4.
36. Morris JS, Jiao D. chemViz2 - Cheminformatics App for Cytoscape. Version 1.1.0 [software]. **2021** [cited 2022 Jan 11]. Available from: https://apps.cytoscape.org/apps/chemviz2.
37. Maitre L, Lau CHE, Vizcaino E, Robinson O, Casas M, Siskos AP*, et al.* Assessment of metabolic phenotypic variability in children's urine using H-1 NMR spectroscopy. Scientific Reports **2017**;7 doi 10.1038/srep46082.
38. Wiedeman AM, Dyer RA, Green TJ, Xu Z, Barr SI, Innis SM*, et al.* Variations in plasma choline and metabolite concentrations in healthy adults. Clin Biochem **2018**;60:77-83 doi 10.1016/j.clinbiochem.2018.08.002.

39. Loef M, von Hegedus JH, Ghorasaini M, Kroon FPB, Giera M, Ioan-Facsinay A*, et al.* Reproducibility of Targeted Lipidome Analyses (Lipidyzer) in Plasma and Erythrocytes over a 6-Week Period. Metabolites **2020**;11(1) doi 10.3390/metabo11010026.
40. Agueusop I, Musholt PB, Klaus B, Hightower K, Kannt A. Short-term variability of the human serum metabolome depending on nutritional and metabolic health status. Sci Rep **2020**;10(1):16310 doi 10.1038/s41598-020-72914-7.
41. Wang Y, Hodge RA, Stevens VL, Hartman TJ, McCullough ML. Identification and Reproducibility of Plasma Metabolomic Biomarkers of Habitual Food Intake in a US Diet Validation Study. Metabolites **2020**;10(10) doi 10.3390/metabo10100382.
42. Wang Y, Hodge RA, Stevens VL, Hartman TJ, McCullough ML. Identification and Reproducibility of Urinary Metabolomic Biomarkers of Habitual Food Intake in a Cross-Sectional Analysis of the Cancer Prevention Study-3 Diet Assessment Sub-Study. Metabolites **2021**;11(4) doi 10.3390/metabo11040248.
43. Li-Gao R, Hughes DA, le Cessie S, De Mutsert R, Den Heijer M, Rosendaal FR*, et al.* Assessment of reproducibility and biological variability of fasting and postprandial plasma metabolite concentrations using 1H NMR spectroscopy. PLoS ONE **2019**;14(6) doi 10.1371/journal.pone.0218549.
44. Breier M, Wahl S, Prehn C, Fugmann M, Ferrari U, Weise M*, et al.* Targeted metabolomics identifies reliable and stable metabolites in human serum and plasma samples. PLoS One **2014**;9(2):e89728 doi 10.1371/journal.pone.0089728.
45. Zheng Y, Yu B, Alexander D, Couper DJ, Boerwinkle E. Medium-term variability of the human serum metabolome in the Atherosclerosis Risk in Communities (ARIC) study. Omics **2014**;18(6):364-73 doi 10.1089/omi.2014.0019.
46. Townsend MK, Clish CB, Kraft P, Wu C, Souza AL, Deik AA*, et al.* Reproducibility of metabolomic profiles among men and women in 2 large cohort studies. Clin Chem **2013**;59(11):1657-67 doi 10.1373/clinchem.2012.199133.
47. Maitre L, Lau CE, Vizcaino E, Robinson O, Casas M, Siskos AP*, et al.* Assessment of metabolic phenotypic variability in children's urine using (1)H NMR spectroscopy. Sci Rep **2017**;7:46082 doi 10.1038/srep46082.
48. Carayol M, Licaj I, Achaintre D, Sacerdote C, Vineis P, Key TJ*, et al.* Reliability of Serum Metabolites over a Two-Year Period: A Targeted Metabolomic Approach in Fasting and Non-Fasting Samples from EPIC. Plos One **2015**;10(8) doi 10.1371/journal.pone.0135437.
49. Xiao Q, Moore SC, Boca SM, Matthews CE, Rothman N, Stolzenberg-Solomon RZ*, et al.* Sources of variability in metabolite measurements from urinary samples. PLoS One **2014**;9(5):e95749 doi 10.1371/journal.pone.0095749.
50. Midttun O, Townsend MK, Nygård O, Tworoger SS, Brennan P, Johansson M*, et al.* Most blood biomarkers related to vitamin status, one-carbon metabolism, and the kynurenine pathway show adequate preanalytical stability and within-person reproducibility to allow assessment of exposure or nutritional status in healthy women and cardiovascular patients. J Nutr **2014**;144(5):784-90 doi 10.3945/jn.113.189738.
51. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N*, et al.* HMDB: the Human Metabolome Database. Nucleic Acids Res **2007**;35(Database issue):D521-6 doi 10.1093/nar/gkl923.
52. Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vázquez-Fresno R*, et al.* HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res **2018**;46(D1):D608-d17 doi 10.1093/nar/gkx1089.
53. Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu Y*, et al.* HMDB 3.0--The Human Metabolome Database in 2013. Nucleic Acids Res **2013**;41(Database issue):D801-7 doi 10.1093/nar/gks1065.
54. Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B*, et al.* HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res **2009**;37(Database issue):D603-10 doi 10.1093/nar/gkn810.

55. Shibata K, Fukuwatari T, Watanabe T, Nishimuta M. Intra- and inter-individual variations of blood and urinary water-soluble vitamins in Japanese young adults consuming a semi-purified diet for 7 days. J Nutr Sci Vitaminol (Tokyo) **2009**;55(6):459-70 doi 10.3177/jnsv.55.459.
56. Wang YX, Feng W, Zeng Q, Sun Y, Wang P, You L*, et al.* Variability of Metal Levels in Spot, First Morning, and 24-Hour Urine Samples over a 3-Month Period in Healthy Adult Chinese Men. Environ Health Perspect **2016**;124(4):468-76 doi 10.1289/ehp.1409551.
57. Chen HG, Chen YJ, Chen C, Tu ZZ, Lu Q, Wu P*, et al.* Reproducibility of essential elements chromium, manganese, iron, zinc and selenium in spot samples, first-morning voids and 24-h collections from healthy adult men. Br J Nutr **2019**;122(3):343-51 doi 10.1017/s0007114519001193.
58. Laenen A, Vangeneugden T, Geys H, Molenberghs G. Generalized reliability estimation using repeated measurements. British Journal of Mathematical and Statistical Psychology **2006**;59(1):113-31 doi https://doi.org/10.1348/000711005X66068.
59. Al-Delaimy WK, Natarajan L, Sun X, Rock CL, Pierce J. Reliability of plasma carotenoid biomarkers and its relation to study power. Epidemiology **2008**;19(2):338-44 doi 10.1097/EDE.0b013e3181635dc2.
60. Williams AE, Maskarinec G, Franke AA, Stanczyk FZ. The temporal reliability of serum estrogens, progesterone, gonadotropins, SHBG and urinary estrogen and progesterone metabolites in premenopausal women. BMC Womens Health **2002**;2(1):13 doi 10.1186/1472-6874-2-13.
61. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med **2016**;15(2):155-63 doi 10.1016/j.jcm.2016.02.012.
62. Andersson A, Marklund M, Diana M, Landberg R. Plasma alkylresorcinol concentrations correlate with whole grain wheat and rye intake and show moderate reproducibility over a 2- to 3-month period in free-living Swedish adults. J Nutr **2011**;141(9):1712-8 doi 10.3945/jn.111.139238.
63. Willett WC. Nutritional epidemiology. Oxford [u.a.]: Oxford University Press; 2013. p. 96-141.
64. Perrier F, Giorgis-Allemand L, Slama R, Philippat C. Within-subject Pooling of Biological Samples to Reduce Exposure Misclassification in Biomarker-based Studies. Epidemiology **2016**;27(3):378-88 doi 10.1097/ede.0000000000000460.
65. Barrera-Gómez J, Spiegelman D, Basagaña X. Optimal combination of number of participants and number of repeated measurements in longitudinal studies with time-varying exposure. Statistics in Medicine **2013**;32(27):4748-62 doi 10.1002/sim.5870.

Table 1 Results from the meta-analysis for nine urinary metabolite concentrations adjusted for creatinine and one unadjusted blood metabolite

Metabolite	**Number of studies**	**ICC (95% CI)**	**I^2**	**Kendall's Tau, p value**
BPA	13	0.15 (0.10, 0.21)	13.9%	0.67, *< 0.001*
MEP	9	0.43 (0.23, 0.63)	75.3%	0.23, 0.37
MnBP	9	0.31 (0.17, 0.46)	75.7%	0.2, 0.45
MEHP	10	0.32 (0.22, 0.42)	29.5%	0.2, 0.3
MEHHP	9	0.20 (0.04, 0.36)	91.9%	0.55, *0.01*
MBzP	9	0.38 (0.24, 0.52)	43.8%	0.38, 0.16
MEOHP	8	0.21 (0.01, 0.40)	97.8%	0.59, *0.01*
Methylparaben	5	0.44 (0.29, 0.59)	71.2%	0.09, 0.76
Propylparaben	5	0.49 (0.32, 0.66)	80.1%	-0.13, 0.65
25(OH)D	7	0.95 (0.90, 0.99)	37.3%	-0.88, *< 0.001*

NOTE: CI, confidence interval.

Figure Legends

Figure 1 Flow diagram of selection process for the inclusion in the qualitative synthesis. The diagram depicts the flow of information throughout the three phases of the systematic literature review. It provides an overview of the number of identified references from the database search and other sources, the number of included references and reasons for exclusion.

Figure 2 Classification of single metabolites (N= 352) according to simplified class membership and visualisation of number of ICC estimates available. All 352 metabolites are grouped by colour into sub-categories according to overall class, which are indicated by the grey square. The lines, so called edges symbolise the chemical similarity between metabolites.

653 The size of the circle, i.e., node, indicates the number of ICC estimates available for the
654 metabolite. These results are based on the extracted data from 178 studies (excluding 14 studies
655 reporting only summary ICC estimates).

656 **Figure 3** Colour coded classification (low to high reproducibility) of mean ICC estimates of
657 single metabolites (N= 352) grouped according to chemical class membership. All 352
658 metabolites are grouped into sub-categories according to overall class, which are indicated by
659 the grey squares. The lines, so called edges symbolise the chemical similarity between
660 metabolites. The colour of the circle, i.e., node, indicates poor (< 0.4) to moderate, good, and
661 excellent (>= 0.85) reproducibility. These results are based on the extracted data from 178
662 studies (excluding 14 studies reporting only summary ICC estimates).

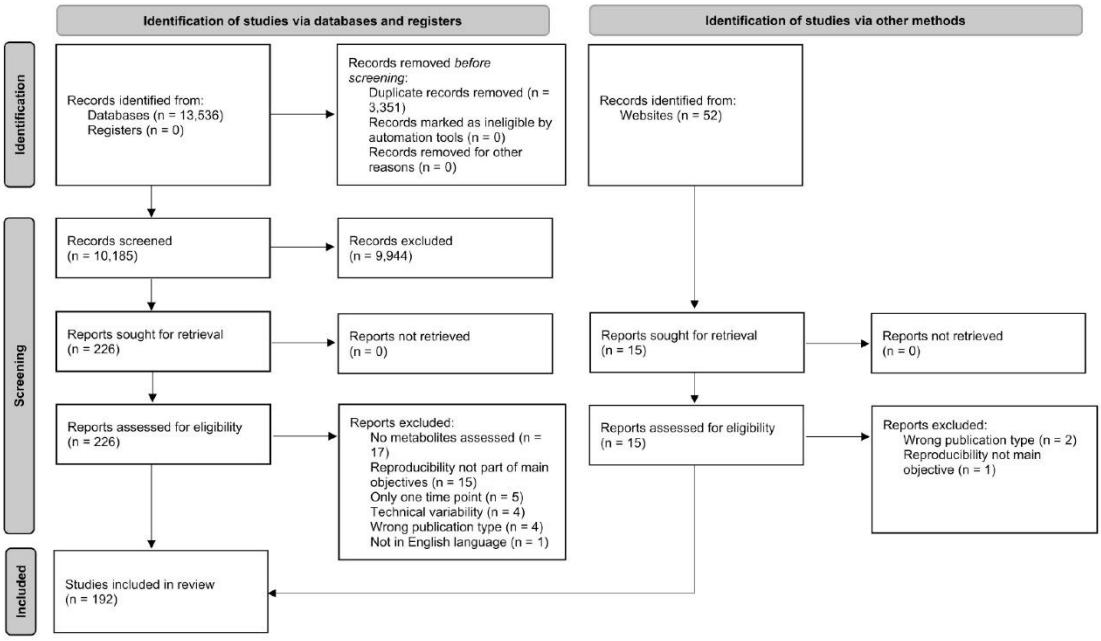
Identification of studies via databases and registers
Identification of studies via other methods
Identification
Records identified from:
Databases (n = 13,536)
Registers (n = 0)
Records removed *before screening*:
Duplicate records removed (n = 3,351)
Records marked as ineligible by automation tools (n = 0)
Records removed for other reasons (n = 0)
Records identified from:
Websites (n = 52)
Screening
Records screened
(n = 10,185)
Records excluded
(n = 9,944)
Reports sought for retrieval
(n = 226)
Reports not retrieved
(n = 0)
Reports sought for retrieval
(n = 15)
Reports not retrieved
(n = 0)
Reports assessed for eligibility
(n = 226)
Reports excluded:
No metabolites assessed (n = 17)
Reproducibility not part of main objectives (n = 15)
Only one time point (n = 5)
Technical variability (n = 4)
Wrong publication type (n = 4)
Not in English language (n = 1)
Reports assessed for eligibility
(n = 15)
Reports excluded:
Wrong publication type (n = 2)
Reproducibility not main objective (n = 1)
Included
Studies included in review
(n = 192)

Figure 1

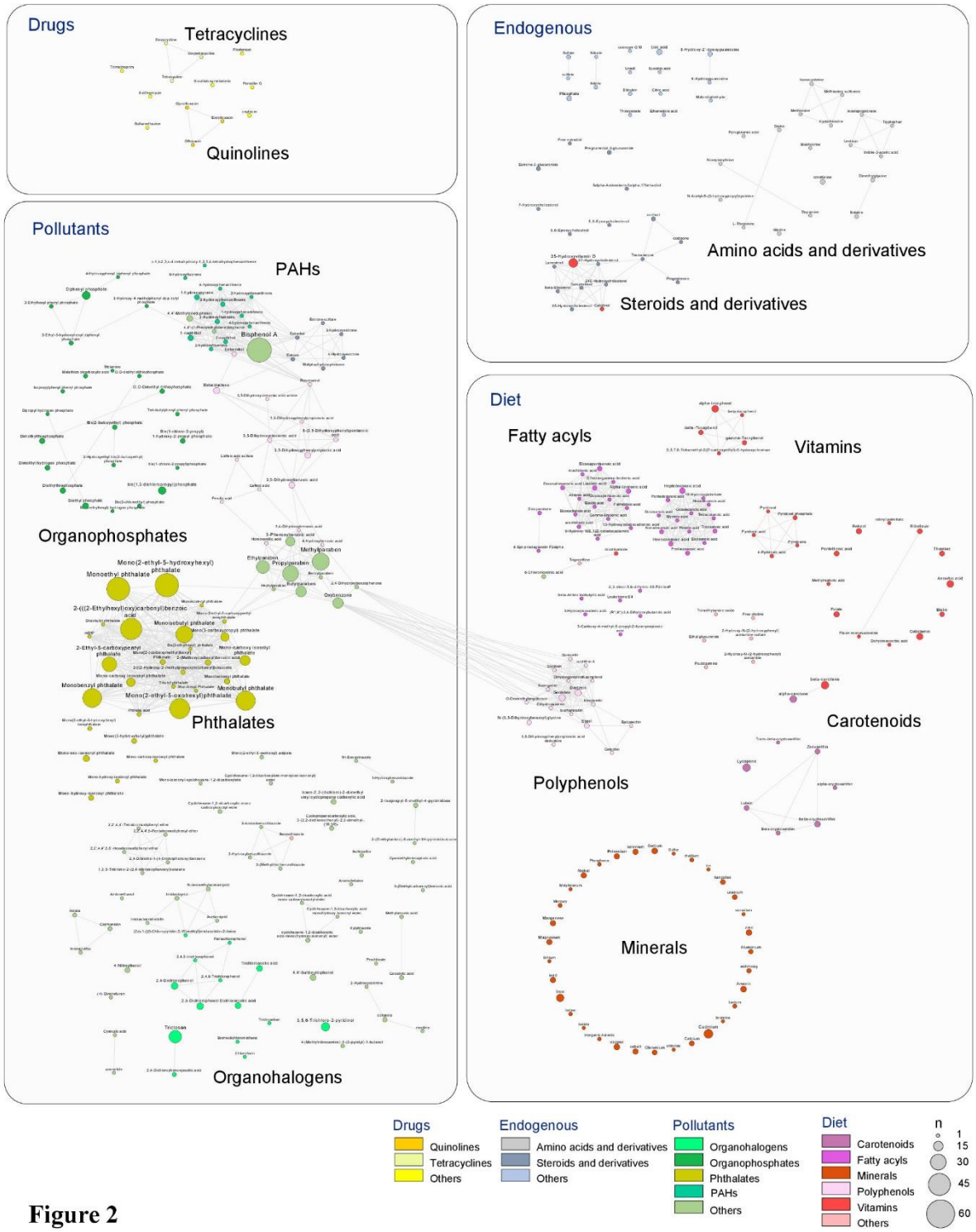
Drugs
Tetracyclines
Quinolines
Endogenous
Amino acids and derivatives
Steroids and derivatives
Pollutants
PAHs
Bisphenol A
Organophosphates
Methylparaben
Ethylparaben
Propylparaben
Benzophenone
Oxybenzone
Mono(2-ethyl-5-hydroxyhexyl) phthalate
Monoethyl phthalate
Monoisobutyl phthalate
Monobenzyl phthalate
Monobutyl phthalate
Mono(2-ethyl-5-oxohexyl)phthalate
Phthalates
Triclosan
Organohalogens
Diet
Fatty acyls
Vitamins
Carotenoids
Polyphenols
Minerals
Drugs
Quinolines
Tetracyclines
Others
Endogenous
Amino acids and derivatives
Steroids and derivatives
Others
Pollutants
Organohalogens
Organophosphates
Phthalates
PAHs
Others
Diet
Carotenoids
Fatty acyls
Minerals
Polyphenols
Vitamins
Others
n
1
15
30
45
60

Figure 2

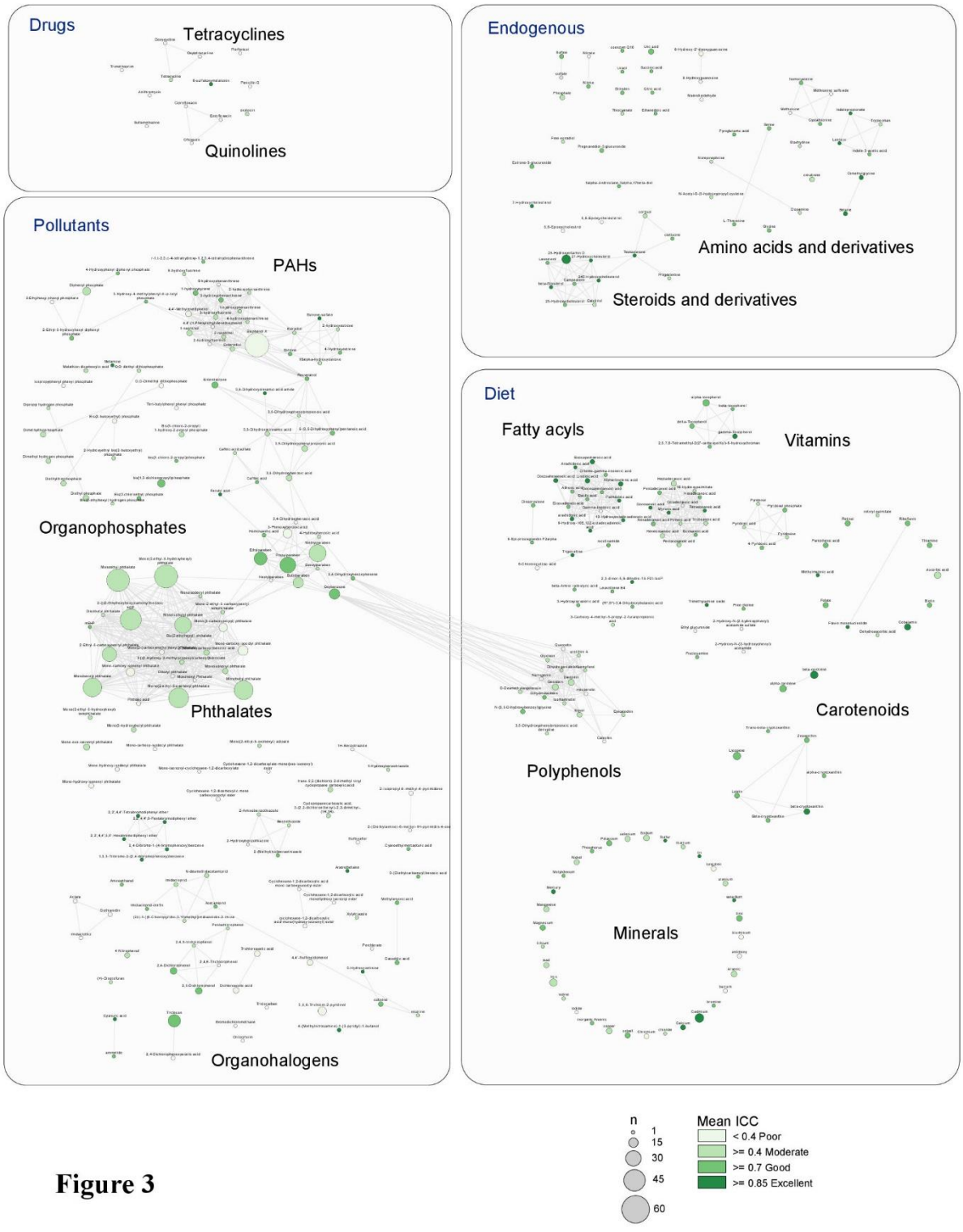
Drugs
Tetracyclines
Quinolines
Endogenous
Amino acids and derivatives
Steroids and derivatives
Pollutants
PAHs
Organophosphates
Phthalates
Organohalogens
Diet
Fatty acyls
Vitamins
Carotenoids
Polyphenols
Minerals
n
1
15
30
45
60
Mean ICC
< 0.4 Poor
>= 0.4 Moderate
>= 0.7 Good
>= 0.85 Excellent

Figure 3