Journal of the American Medical Informatics Association, 25(11), 2018, 1432–1443
doi: 10.1093/jamia/ocy119
Advance Access Publication Date: 26 October 2018
Research and Applications

Research and Applications

Development and validation of the PEPPER framework (Prenatal Exposure PubMed ParsER) with applications to food additives

Mary Regina Boland,[1,2,3,4] Aditya Kashyap,[5] Jiadi Xiong,[5] John Holmes,[1,2] and Scott Lorch[6]

[1]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA, [2]Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, USA, [3]Center for Excellence in Environmental Toxicology, University of Pennsylvania, Philadelphia, PA, USA, [4]Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA, USA, [5]Data Science Masters Program, University of Pennsylvania, Philadelphia, PA, USA and [6]Division of Neonatology, Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, PA, USA

Corresponding Author: Mary Regina Boland, 423 Guardian Drive, 421 Blockley Hall, Philadelphia, PA 19104, USA (bolandm@upenn.edu)

Received 18 May 2018; Revised 20 July 2018; Editorial Decision 9 August 2018; Accepted 13 August 2018

ABSTRACT

Background: Globally, 36% of deaths among children can be attributed to environmental factors. However, no comprehensive list of environmental exposures exists. We seek to address this gap by developing a literature-mining algorithm to catalog prenatal environmental exposures.

Methods: We designed a framework called **PEPPER**: **P**renatal **E**xposure **P**ubMed **P**ars**ER** to a) catalog prenatal exposures studied in the literature and b) identify study type. Using PubMed Central, PEPPER classifies article type (methodology, systematic review) and catalogs prenatal exposures. We coupled PEPPER with the FDA's food additive database to form a master set of exposures.

Results: We found that of 31 764 prenatal exposure studies only 53.0% were methodology studies. PEPPER consists of 219 prenatal exposures, including a common set of 43 exposures. PEPPER captured prenatal exposures from 56.4% of methodology studies (9492/16 832 studies). Two raters independently reviewed 50 randomly selected articles and annotated presence of exposures and study methodology type. Error rates for PEPPER's exposure assignment ranged from 0.56% to 1.30% depending on the rater. Evaluation of the study type assignment showed agreement ranging from 96% to 100% (kappa $= 0.909$, $p < .001$). Using a gold-standard set of relevant prenatal exposure studies, PEPPER achieved a recall of 94.4%.

Conclusions: Using curated exposures and food additives; PEPPER provides the first comprehensive list of 219 prenatal exposures studied in methodology papers. On average, 1.45 exposures were investigated per study. PEPPER successfully distinguished article type for all prenatal studies allowing literature gaps to be easily identified.

Key words: prenatal exposure, food additives, literature mining

INTRODUCTION

Importance of environment in human health and during the prenatal period

Environmental exposures are critically important for understanding human health and disease. Globally, 23% of all deaths and 36% of deaths among children (ages 0-14 years) can be attributed to environmental factors.[1] Several types of factors are collectively termed "environment" by researchers. These include **lifestyle** factors (exercise, diet, and stress), **climate** factors (sunlight, precipitation, and wind speed), **socioeconomic** factors (occupation, income, insurance status), **pharmacological** factors (opioids, antidepressants, Non-Steroidal Anti-Inflammatory Drugs or NSAIDS) and **pollutant** factors (phthalates, fine and coarse air particulates).[2] These 5 types of "factors" are collectively termed "environmental exposures" by the scientific community. Each of these plays a role in variance in health outcomes. However, **no** comprehensive list of environmental exposures exists.

Informatics and exposome-related science

Several informatics methods have investigated environmental perturbations and their effects on human health and disease. Studies have used Electronic Health Record (EHR) data to investigate environmental challenge in disease progression,[3] identified thyroid cancer hotspots in Vermont,[4] and correlated air pollution with disease risk in Italy.[5] Environment-Wide Association Studies have been performed using EHR data.[6] Boland et al. developed a method correlating birth season[7,8] and trimester information with climate and pollution variables.[2] Boland et al demonstrated that informatics methods can overcome EHR biases[9,10] and be used to probe the disease-environment interaction[2] with findings confirmed in canines.[11] There remains a need for informatics methods to investigate environmental effects on human health following prenatal exposure.

Importantly, many studies investigating the effect of the environment on human disease often look at a **single** environmental factor (eg air pollution) without exploring simultaneous exposure to multiple environmental exposures. The classic EWAS study investigated relationships between all factors measured in the National Health and Nutrition Examination Survey (NHANES) and their relationship with 1 outcome - type 2 diabetes.[6] However, the NHANES contains elements from a survey and is not a comprehensive list of all environmental exposures. This purpose of this study is to derive a comprehensive list of environmental exposures from the environmental exposures studied in biomedical literature.

Literature mining of PubMed Central and cataloging exposures

Many methods utilize publically available abstracts from PubMed[12] including PubMatrix,[13] PubTator,[14] LitInspector,[15] PolySearch[16] and COSMIC.[17] Some methods use PubMed abstracts as 1 source in their biological networks, such as the STRING database.[18] Another method links genes, diseases and drugs together into biological networks[19] while another method linked PubMed abstracts with information from EHRs for pancreatic cancer prediction.[20] Many methods use PubMed abstracts exclusively because the full text remains behind a pay wall. These articles often list exclusive use of article abstracts as a critical limitation.[16] Recently, a study demonstrated that using the full text outperformed use of only abstracts.[21]

Alternatively, PubMed Central (PMC) is a subset of PubMed containing only freely accessible full-text manuscripts. Using the full text of the manuscript allows an algorithm to be constructed that distinguishes methodology type (eg original research articles, non-systematic reviews, editorials, perspectives, and clinical practice guidelines). Study type is important because some prenatal exposures are discussed frequently in the literature in guidelines and perspectives, but rarely studied in original research articles. Therefore, a gap in the literature exists for those exposures that would not be readily apparent from a quick search in Pubmed. We constructed our algorithm to identify study methodology type to easily identify literature gaps for environmental exposures.

The purpose of this study is to develop a methodological framework that uses knowledge from the literature to derive a comprehensive list of environmental exposures studied during the prenatal period. This will allow researchers to assess the "state-of-the-field" while identifying literature gaps in the prenatal exposure space. In addition, we provide the first list of environmental exposures studied during the prenatal period for other researchers to use. The method we developed is called **PEPPER: P**renatal **E**xposure **P**ubMed **P**ars**ER. PEPPER** uses full-text research articles from PMC to identify environmental exposures studied during the prenatal period in methodological studies. Using PEPPER, we are able to identify exposures that are studied more frequently in review papers vs methodological studies, which may indicate a form of publication bias.[22–25] We provide a case study of PEPPER's use within the food additive domain and identify literature gaps, which can be addressed by further study.

METHODS

Dataset: Pubmed Central

PubMed Central (https://www.ncbi.nlm.nih.gov/pmc/) is a publicly available database maintained by the National Library of Medicine that provides the full text of each article. We used PMC as our literature data source for this study. PEPPER used the *requests* and *BeautifulSoup* libraries of Python 2.7, along with PMCID lists obtained from PMC.

Preliminary steps: extracting all relevant prenatal exposure articles

A simple query of "prenatal" and "exposure" on PMC retrieves 48 066 articles (query performed in March 2018). However, we found that many of those articles mentioned the term "prenatal" and "exposure" in the reference section making the results not relevant. This was especially true for environmental exposures. For example, general studies on lead, including chemical synthesis and reactivity studies would cite a reference on prenatal exposure to lead because of the media attention placed on the importance of prenatal exposure to lead. We refined the query and restricted the terms "prenatal" and "exposure" to the body of the manuscript excluding the references using the query: `("prenatal"[Body—All Words] AND "exposure"[Body—All Words])`. Additional details are available in Supplementary Appendix. This resulted in relevant retrieved articles and a dataset of 31 764 studies.

Construction of PEPPER framework

We extracted all prenatal exposure papers with mention of the terms prenatal and exposure in the body of the text (excluding reference mentions). The PEPPER framework a) identifies articles that are methodological in nature and b) extracts relevant exposures identified in those studies. We designed PEPPER to identify whether or not a study was methodological because no "method" tag exists for

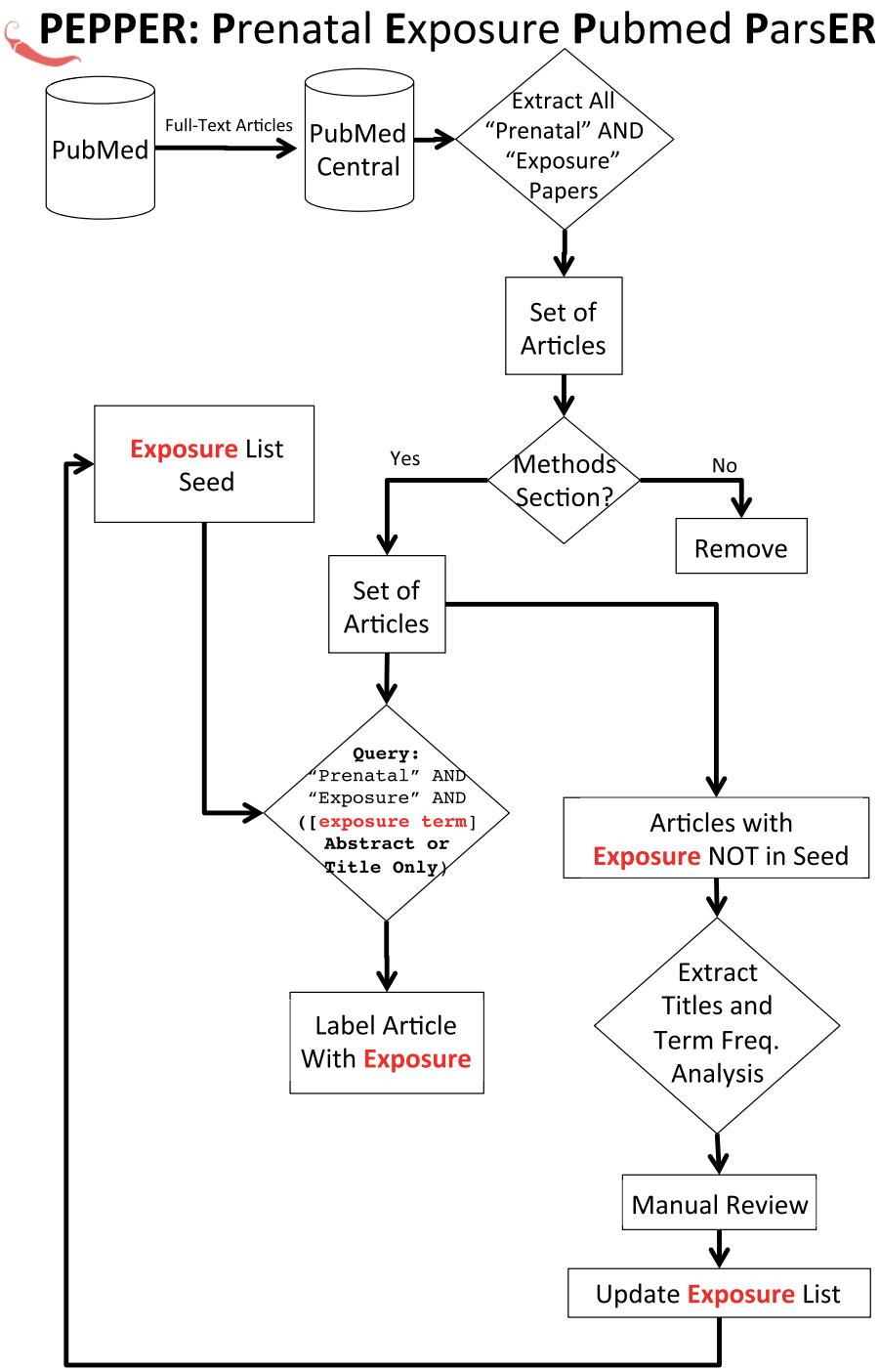

Figure 1. PEPPER flowchart detailing the steps of the algorithmic framework.

searching PMC. The corpus of 31 764 articles served as input into our PEPPER Framework along with an initial exposure list seed (development described subsequently, Figure 1).

The first step was to query PMC to retrieve articles relevant for specific exposures. An example of the PMC query is shown below. We employ the Arden syntax "curly bracket" notation[26] to represent the exposure variables that PEPPER iterates over:

```
("prenatal"[Body–All Words] AND "exposure"[Body–All Words]) AND ("{term}"[Title] OR "{term}"[Abstract])
```

where "{**term**}" is the particular chemical or contaminant of interest.

We decided to restrict the exposure **term** to occur only within the abstract and title because certain exposures (eg "phthalates") were mentioned frequently in the Introduction and Background of prenatal exposure papers (even if the topic of the paper was about a different exposure, such as "mercury"). We made one exception for lead. We used the chemical element name "pb" (which includes "Pb") instead of "lead" because "lead" is used commonly in other contexts (eg as a verb). PEPPER ignores capitalization of terms.

Distinguishing study methodology type

We used the following steps to identify articles with methods from those without methods. **Step 1**, we acquired all h2 section headers for each article. **Step 2**, the html was converted to lxml markup. **Step 3**, we searched for headings containing the term "method" in the h2 header (eg "Materials and Methods," "Methodological Framework").

To distinguish methodological review papers (eg systematic reviews) from other methodological papers, we searched for the word "review" in the title. This allowed us to distinguish papers without methods, review papers with a methods section, and non-review papers with a methods section (ie the "original research studies"). Importantly, we only used the word "review" in the title of the paper for methodology studies to separate out the systematic reviews from original research articles because both article types have methods sections. Reviews without methods sections were binned in the "non-methodological study" category.

Development of initial exposure seed list and iterative refinement

To retrieve articles related to specific exposures, we needed an initial exposure list seed. We carefully reviewed all articles published between April 2016 and September 2017 and manually extracted exposures from methodology studies and meta-analyses. This resulted in a set of 419 articles, which included reviews and other non-methodology studies. Our manual review resulted in an initial exposure list of 29 exposures. Another member of the team carefully reviewed this list, and both individuals met to discuss any discrepancies.

PEPPER semi-automatic exposure generation process

Each exposure term from the initial 29-exposure seed was queried iteratively (one exposure at a time). Querying PMC for "prenatal" and "exposure" returned some articles that did not study an exposure listed in the initial seed set. We retained these articles and then performed additional processing to extract titles and calculate term frequencies. Titles were chosen because the majority of articles can be understood using the title alone.[27] We extracted all titles from the articles **not** in the exposure list. The scraped articles (in html) were processed in the following manner. First, stop words were removed using the *nltk* library of python. Next, using the *re* library of python, regular expressions were written to remove all the content within brackets, along with figure numbers and table numbers. Words such as "he/she" and "male/female" that contained a backslash within it or any other special character were also removed. Words which were 3 letters or shorter in length were removed in order to eliminate abbreviations. Finally, we eliminated non-ASCII characters, such as "ä" and "æ." PEPPER also ignores the case (ie capitalization) of the terms. Term frequency analysis on these titles allowed for easy manual review and subsequent updating of the environmental exposure seed list. The term frequency lists were manually reviewed by two co-authors (AK, MRB) to ensure that non-environmental exposure topics, including age, schooling, and so forth were removed. We coupled the 43-exposure set with a database of food additives, described in section 2.5 for a final exposure set of 219 exposures (Supplementary Files S1 and S2).

PEPPER exposure extraction and labeling across entire corpus

PEPPER extracted and labeled all articles in the 31 764-article corpus with the appropriate exposures. Studies were not assigned a particular exposure by PEPPER until the entire set of exposures was defined. In addition a study could be assigned multiple exposures given that some studies investigate multiple exposures. For example, 1 study may investigate the effects of obesity, high-fat diet and smoking on prenatal development. In that case "obesity," "high-fat diet," and "smoking" would each be assigned to the study.

Sex-specific study analysis

Studies were defined as having sex-specific outcomes, if they were returned following a separate PMC query given below:

```
("prenatal"[Body—All Words] AND "exposure"[Body—
All Words]) AND ("{term}"[Title] OR
"{term}"[Abstract]) AND (("sex specific"[Title]
OR "sex differences"[Title]) OR ("sex specific"
[Abstract] OR "sex differences" [Abstract]))
```

where "**{term}**" is the particular chemical or contaminant of interest.

An example query for "alcohol" is given below:

```
("prenatal"[Body—All Words] AND "exposure"[Body—
All Words]) AND ("alcohol"[Title] OR
"alcohol"[Abstract]) AND (("sex specific"[Title]
OR "sex differences"[Title]) OR ("sex specific"
[Abstract] OR "sex differences"[Abstract]))
```

Knowledge of sex-specific outcomes is important for exposures, especially Endocrine Disrupting Chemicals (EDCs). Sex-specific outcomes were only explored in the 43 common exposures.

Methods for an example case study for PEPPER: food additives

Food additives have been studied since the 1940s[28] and continue to be studied throughout the following decades.[29,30] For our case study, we expanded our exposure list to include all possible food additives. The United States Food and Drug Administration (FDA) maintains a database of food additives called "Everything Added to Food in the United States" or EAFUS. EAFUS contains information on 3968 compounds. We downloaded the publicly available EAFUS, including information on all 3968 food additives in February 2018 (https://www.fda.gov/Food/IngredientsPackagingLabeling/FoodAdditivesIngredients/ucm115326.htm). We used only publicly available information and did not clean this list, but used the raw terms in our case study of PEPPER.

As shown in Table 1, each compound is assigned a toxicological status code. There are 6 categories provided (Table 1), we grouped these into 4 categories for ease of interpretation. We provide our 4 super-categories - 1) commonly used; 2) new; 3) limited to no use; and 4) banned -along with the original EAFUS group in Table 1. All 3968 food additive compounds were run through PEPPER. We only made one small modification to PEPPER removing the quotes around the exposure **term.** By removing the quotes around the exposure **term**, we are making use of PMC's built-in term expansion software, which is critical given that these exposures were not manually curated. For example, the food additive "calamus root," would be queried as follows:

```
("prenatal"[Body—All Words] AND "exposure"[Body—
All Words]) AND (calamus root[Title] OR calamus
root[Abstract])
```

Table 1. EAFUS food additive categories and PEPPER super-categories

EAFUS Food Additive Categories		
Code	Category Description	PEPPER Super- Category
ASP	Fully up-to-date toxicology information has been sought.	Commonly_used
EAF	There is reported use of the substance, but it has not yet been assigned for toxicology literature search.	New
NEW	There is reported use of the substance, and an initial toxicology literature search is in progress.	New
NIL	Although listed as added to food, there is no current reported use of the substance, and, therefore, although toxicology information may be available in PAFA, it is not being updated.	Limited_use
NUL	There is no reported use of the substance and there is no toxicology information available in PAFA.	Limited_use
BAN	The substance was formerly approved as a food additive but is now banned; there may be some toxicology data available.	Banned

Evaluation of PEPPER

Our evaluation for PEPPER consisted of 3 parts: 1) evaluation of the exposure assignment; 2) evaluation of the study type assignment, and 3) evaluation of the recall of PEPPER. For the first part, 2 raters manually reviewed 50 retrieved articles by PEPPER and annotated exposures using a set of 43 exposures (not including EAFUS). Because each article could have multiple exposures, each rater annotated all exposures mentioned in the title and abstract. For the evaluation of the study type assignment, another random set of 50 articles was extracted. Each rater assigned the manuscript as a methodology paper, non-methodology paper, PDF-only paper or systematic review. We distinguish PDF papers from other types of papers, because PEPPER (along with PMC's search functionality) cannot parse the full text of PDF-only studies. After each rater conducted their evaluations independently, they met to discuss differences and develop consensus.

Evaluating the recall of PEPPER

In information retrieval, recall is assessed to determine how many of a set of gold standard relevant documents are retrieved by a given algorithm. Our gold-standard consists of a systematic review and meta-analysis published in 2017 on fetal alcohol spectrum disorder.[31] This paper contained 86 references that are key in the field. References were excluded from the gold-standard if they were: a) websites; b) software packages (eg python, R); c) statistical methods papers, and d) references to background birth rates that were not exposure-related. This process resulted in a set of 64 gold-standard relevant alcohol exposure studies. We reviewed each study for presence/absence in PMC and if absent from PMC then presence/absence in Pubmed was ascertained. Of 64 relevant articles, only 18 were found in PMC. PEPPER's recall was determined using these 18 relevant alcohol exposure studies.

RESULTS

For researchers interested in running an article or a set of articles through PEPPER, we provide code and relevant datasets at: https://github.com/bolandlab/PEPPER/.

Corpus

Our final prenatal exposure study corpus consisted of all articles on PMC with a mention of "prenatal" and "exposure" within the body of the article. The final set included 31 764 articles ranging in publication date from 1921 through 2018.

Study methodology type varies by prenatal exposure

We investigated the study methodology type and how it varies by prenatal exposure. Overall, 16 832 prenatal exposure articles were methodological out of 31 764 articles (53.0%). We highlight the percent of studies that are methodological and the percent of studies constituting systematic reviews in Figure 2. We sorted exposures by the total number of studies per exposure for clarity. Notice that some exposures, eg iodine (second-highest peak in Figure 2B), have relatively fewer methodological studies and relatively more systematic reviews. This indicates that fewer original research articles contribute to those systematic reviews. Other exposures, such as fluoride (highest peak in Figure 2B) have a relatively large proportion of systematic reviews with 31 total studies including 16 methodological and 3 systematic reviews. Not all exposures that were low in methodological studies were necessarily high in systematic reviews because some non-methodological studies are editorials, perspectives, case studies and clinical practice guidelines.

PEPPER allowed us to compute the number of methodology studies per number of systematic reviews for each exposure. This enabled us to estimate the average number of methodology studies underlying each systematic review. The overall average across all 43 common exposures (excluding exposures with 0 systematic reviews) was 28.95 methodology studies per systematic review with a standard deviation of 28.24. We provide the proportion of methodology studies per systematic review for each exposure in Table 2. Some exposures, such as testosterone, have a high proportion—148—indicating that a large number of methodology studies exist per systematic review. This occurs when an exposure is well studied in the literature. Other exposures, such as HIV and soy, have low proportions. An exposure can have a low proportion for two reasons: 1) there is a relatively large number of systematic reviews given the total number of studies (eg soy) or 2) there is a relatively low number of methodological studies (eg HIV). Exposures such as HIV tend to have a large number of non-methodological studies, including guidelines, case reports and perspectives. We provide the breakdown of study methodology type for each of the 43 common exposures and the 176 EAFUS food exposures as Supplementary files (Supplementary Files S3 and S4).

Studies evaluating sex-specific effects

The percent of prenatal exposure studies investigating sex-specific effects is shown in Figure 3, along with the breakdown of sex-specific methodology studies. Sex-specific variance in exposure outcomes was studied frequently for testosterone, estrogen, endocrine disrupting chemicals (or EDCs), BPA (bisphenol A), high-fat diet, phthalate, and soy.

Studies on exposures such as thiamine, iodine, HIV, vitamin D, malaria, acetaminophen, and radiation did not typically investigate sex-specific outcomes following the exposure. This represents a potential literature gap, and a possibility for measurement bias (only known endocrine disruptors were investigated for a sex-related effect). Therefore, non-endocrine disruptors (eg mercury) were not frequently studied with regards to their sex-related outcomes. This is an example of measurement bias because the outcome is mainly studied among exposures where an effect is expected. The effects of

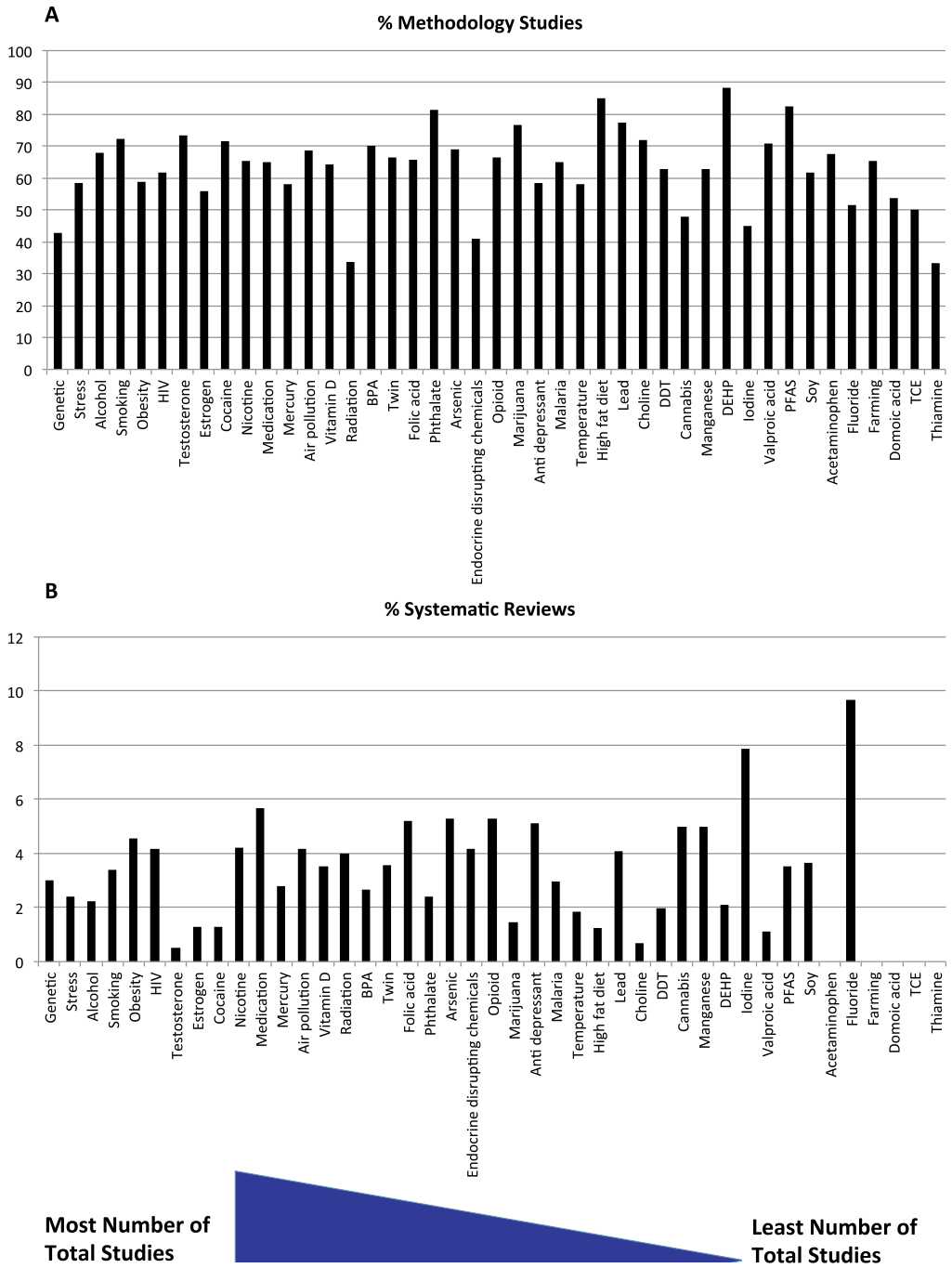

Figure 2. Percent of studies that are methodological vs systematic reviews by prenatal exposure. Exposures are sorted by the total number of studies per exposure. Notice that some exposures have less then 50% of the research consisting of methodological studies, for example radiation, genetics, iodine, and thiamine (Figure 2A). Other exposures are frequently studied in systematic reviews, for example fluoride, iodine, and medications (Figure 2B). Notice that not all exposures that are low in methodological studies are necessarily high in systematic reviews (although some like iodine are), and this is because some non-methodological studies are editorials, perspectives, case studies, and clinical practice guidelines.

pollutants on sex ratios is not fully known and, therefore, worth investigating more broadly (ie not just among known endocrine disrupting chemicals).

Methodological studies investigate multiple exposures

Environmental exposures rarely occur in isolation. We found that on average methodological studies investigated 1.45 prenatal exposures per study (min = 1, max = 8) among the 9492 studies where PEPPER captured at least 1 exposure. To illustrate the high-degree of overlap among studies, we performed Multi-Dimensional Scaling (MDS). MDS is a form of dimensionality reduction that visualizes the overlap of exposures across studies. We performed MDS for all 9492 studies with exposures captured by PEPPER. We grouped exposures to illustrate the breakdown of lifestyle exposures, endocrine disrupting pollutants, non-endocrine disrupting

Table 2. Proportion of methodology studies per systematic review by prenatal exposure

Exposure	Original Research Articles (ie Methodological Studies)	Systematic Review Studies	Num. Methodological Studies / Num. of Systematic Reviews
Genetic	1310	92	14.239
Stress	1600	66	24.242
Alcohol	1269	42	30.214
Smoking	1261	59	21.373
Obesity	737	57	12.93
HIV	488	33	14.788
Testosterone	438	3	146
Estrogen	263	6	43.833
Cocaine	334	6	55.667
Nicotine	293	19	15.421
Medication	262	23	11.391
Mercury	228	11	20.727
Air pollution	263	16	16.438
Vitamin D	237	13	18.231
Radiation	118	14	8.429
BPA	237	9	26.333
Twin	167	9	18.556
Folic acid	164	13	12.615
Phthalate	202	6	33.667
Arsenic	169	13	13
Endocrine disrupting chemicals	89	9	9.889
Opioid	138	11	12.545
Marijuana	157	3	52.333
Anti depressant	103	9	11.444
Malaria	109	5	21.8
Temperature	94	3	31.333
High fat diet	136	2	68
Lead	114	6	19
Choline	105	1	105
DDT	64	2	32
Cannabis	48	5	9.6
Manganese	63	5	12.6
DEHP	84	2	42
Iodine	40	7	5.714
Valproic acid	63	1	63
PFAS	47	2	23.5
Soy	34	2	17
Acetaminophen	25	0	NA*
Fluoride	16	3	5.333
Farming	15	0	NA*
Domoic acid	7	0	NA*
TCE	6	0	NA*
Thiamine	2	0	NA*

*Because there are 0 Systematic Reviews this number is not possible to compute.

pollutants, vitamins and minerals, infections, genetics, traditional environmental exposures (ie air pollution, radiation, temperature), and food additives from EAFUS. Figure 4 illustrates how certain exposure types are studied in patterns that are distinct from other exposure types. For instance, food additive studies are located in regions II and IV (Figure 4) while genetics; infection, non-endocrine disrupting pollutants, smoking/nicotine, and alcohol are located only in region I. Certain exposures are located throughout the graph including the traditional environment exposures, medications and illicit drugs, hormones and endocrine-disrupting pollutants, stress/diet, and vitamins and minerals (Figure 4).

Example case study for PEPPER: food additives

We used PEPPER for a case study on food additives. Using FDA's EAFUS, PEPPER determined how many studies on food additives investigate prenatal exposure. We compared this against the total number of studies mentioning the food additive to obtain the proportion of studies assessing a prenatal effect.

Prenatal exposure effects of food additives were studied for 176 compounds out of 3968 (4.4%) compounds contained in EAFUS. Of 16 832 prenatal exposure methodology studies, only 1886 (11.2%) investigate food additive effects. Among these, 768 (4.6%) investigated only a food additive contained in EAFUS (ie not any of the 43 common exposures) while another 1118 (6.6%) investigated food additives and a commonly studied prenatal exposure (Figure 5).

In total, 3117 studies investigated prenatal exposure to food additives. The majority of these were methodology studies (60.5%), followed by non-methodology studies (27.2%), PDF only (8.9%) and systematic reviews (3.4%) (Table 3). Prenatal exposure to commonly used food additives (EAFUS category ASP) are rarely studied with a rate of only 0.24% of methodology studies. Surprisingly, there is also a paucity of research on the effects of banned food additives on prenatal development. Of 2105 research articles investigating banned food additives, only 4 (0.19%) investigate effects during the prenatal period and only 3 (0.14%) were methodology studies (Table 4).

Evaluation of PEPPER

Two raters independently assigned exposures to 50 articles. PEPPER's error rate for exposure assignment ranged from 0.56% to 1.30% depending on the rater. Evaluation of the study type assignment showed agreement ranging from 96% to 100% depending on the rater (kappa = 0.916, $p < .001$). There was 100% agreement with regards to the PDF-only categorization among both raters and PEPPER. Two discrepancies existed between raters. One involved a state-of-the-art review article on a scientific method, and; therefore, one rater annotated it as a non-methodological paper and the other as a methodological paper. The second discrepancy involved a publication of a set of abstracts from a conference (assigned to one PMC ID). One rater decided this was not a methodological study because it was a set of abstracts (which is the same as PEPPER's assignment) and the other decided that the term "method" was used in the abstracts, and; therefore, this should be considered a methodology study. Overall, agreement was high between raters (Fleiss kappa = 0.916, $p < .001$).

PEPPER's recall

Using a gold standard set of 64 prenatal alcohol exposure studies,[31] we computed PEPPER's recall. Only 18 of the 64 relevant studies were found in PMC (28.1%), most likely this is because PMC requires free access to the manuscript's fulltext. PEPPER successfully retrieved 17 of the 18 relevant studies (94.4%) available within PMC. The only relevant article that PEPPER failed to retrieve was a fetal alcohol syndrome study where the prenatal exposure to alcohol was implied implicitly but not stated explicitly. Diseases and disorders that are based on a prenatal exposure, such as "fetal alcohol

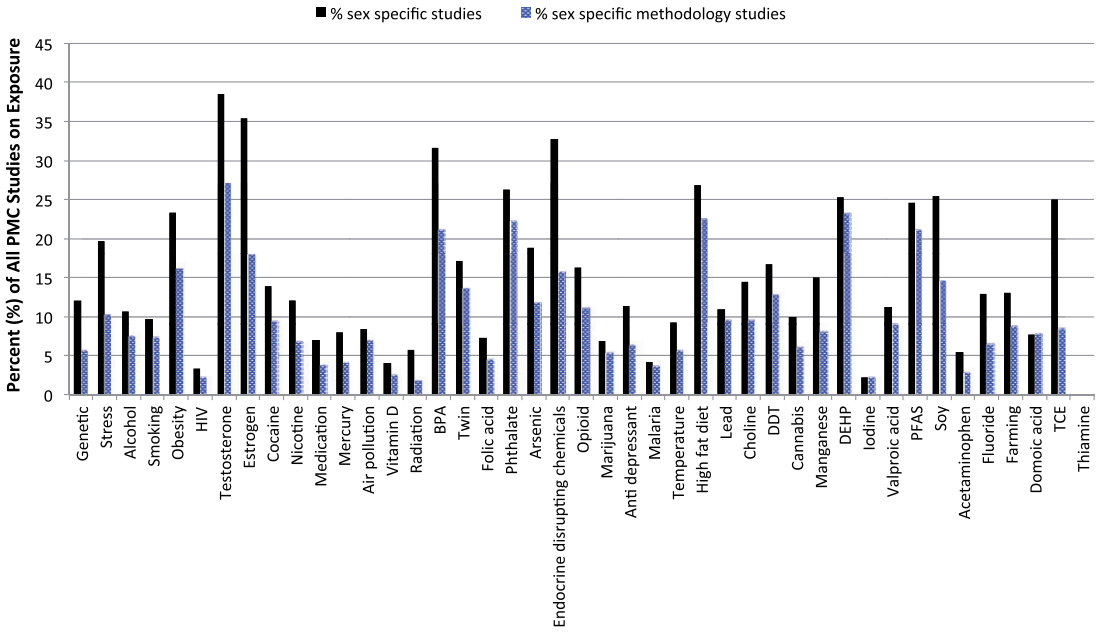

Figure 3. Percent of prenatal exposure studies on a particular exposure that investigates a sex-specific difference. Exposures are sorted by the total number of studies per exposure (far left is the most number of studies, far right is the least number of studies). The percent of all studies that investigate sex-specific effects is shown along with the percent of methodology studies investigating sex-specific effects (out of all studies).

syndrome disorder" sometimes omit mentioning that the exposure was prenatal because this knowledge is implied in the diagnosis.

DISCUSSION

In this study, we describe the development of **PEPPER**: **P**renatal **E**xposure **P**ubMed **P**ars**ER**. PEPPER performs 2 tasks: a) catalogs prenatal exposures studied in the literature and b) identifies study type. PEPPER allowed us to develop a list of commonly studied exposures. We are also able to identify exposures that have a large number of systematic reviews vs methodology studies and exposures that are discuss more in non-methodological studies (eg perspectives, guidelines, and editorials). This allowed us to identify gaps in the literature, especially pertaining to food additives (Figure 5).

A main contribution of PEPPER is the ability to distinguish original research studies (ie methodology studies) from systematic reviews (containing a methods section) and also from non-methodology studies (eg perspectives, editorials, clinical practice guidelines). As of June 2018, PMC has added a "Methods – Key Terms" item to their metadata options for querying, which should facilitate this for other researchers in the future (however as of July 2018, it is still under development and not fully functional). The literature gaps that PEPPER identifies are vital for distinguishing exposures mentioned frequently in non-methodological papers vs commonly reviewed vs studied in formal research articles.

Exposure assignment can be challenging. PEPPER does not distinguish between the main exposure outcome and the confounder variable in a study. Instead, it identifies all exposures present in the title or abstract of the study. Of 9492 articles where we identified an exposure, we found that on average 1.45 exposures were studied per methods paper. In Figure 4, we display the articles by exposure type and the patterns of overlap across studies. Exposures can be grouped by their patterns. Interestingly, food additives formed a unique pattern. Exposures can be grouped by their type (eg lifestyle vs genetics) or by the pattern obtained in Figure 4 illustrating how they are studied. These patterns can be utilized by others studying environmental exposures.

Food additives are ubiquitous in modern day life.[32] For example, caffeine is a well-known food additive with known prenatal effects. Furthermore, it is reported that 85% of the US population ingests a caffeine beverage per day with some estimates up to 89%.[33,34] PEPPER identified 81 studies investigating prenatal exposure effects of caffeine. Therefore, this is a well-studied exposure.

However, across all food additives, PEPPER revealed that less then 1% of studies for each of the major food additive groupings by the FDA investigate prenatal exposure. The highest proportion of prenatal exposure studies was found in the "new" food additives. Additionally, only 60.5% of food additive studies investigating prenatal exposure were methodological studies (Table 3). Across all major food additive categories, >1% of prenatal exposure studies are methodological in nature (Figure 5) illustrating an important literature gap.

Food additives are important not only because of their potential teratogenicity, but also because of food allergies and their effects on the developing fetus. Food allergies are becoming more common with estimates between 1 to 10% of the general population.[35] Since immune reactions during pregnancy can affect fetal development, studying food additives during the prenatal period is important to distinguish the food additive effect from the allergy.

Another important literature gap is the paucity of literature on currently banned food additives. Only 3 articles describing prenatal exposure effects were methodological studies. Only 2 food additives were tested – coumarin and thiourea. All 3 studies were

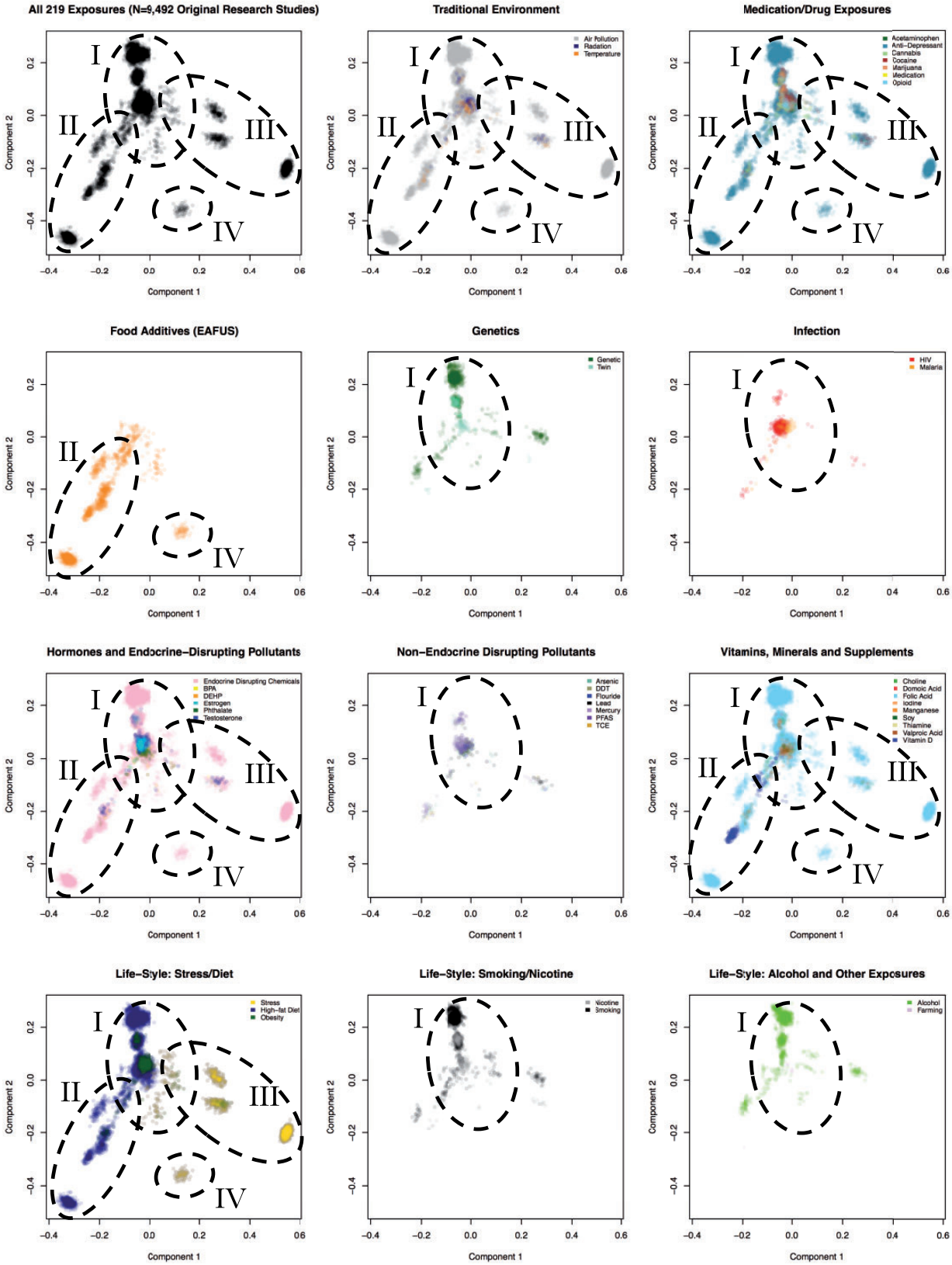

Figure 4. Multi-Dimensional scaling plots visualize the overlap among prenatal exposures present in original research articles. The graph showing all 219 exposures is the upper left hand subplot. For the purpose of the MDS, all 176 EAFUS food additive exposures were grouped together into a single term. Therefore 44 distinct exposures are shown in these subplots. The graphs are divided into 4 keys areas denoted by Roman Numerals (I, II, III, IV). In addition, exposures are grouped according to their types: traditional environment (ie air pollution, radiation, temperature), medications and illicit drugs, food additives from EAFUS, genetics, infection, hormonal disrupting pollutant, non-hormonal disrupting pollutant, vitamins and minerals, and various lifestyle exposures including stress/diet, smoking/nicotine and alcohol and other lifestyle exposures. The overall average number of exposures per study was 1.45 with a range from 1 to 6. The overlap is also easily identified in these MDS subplots.

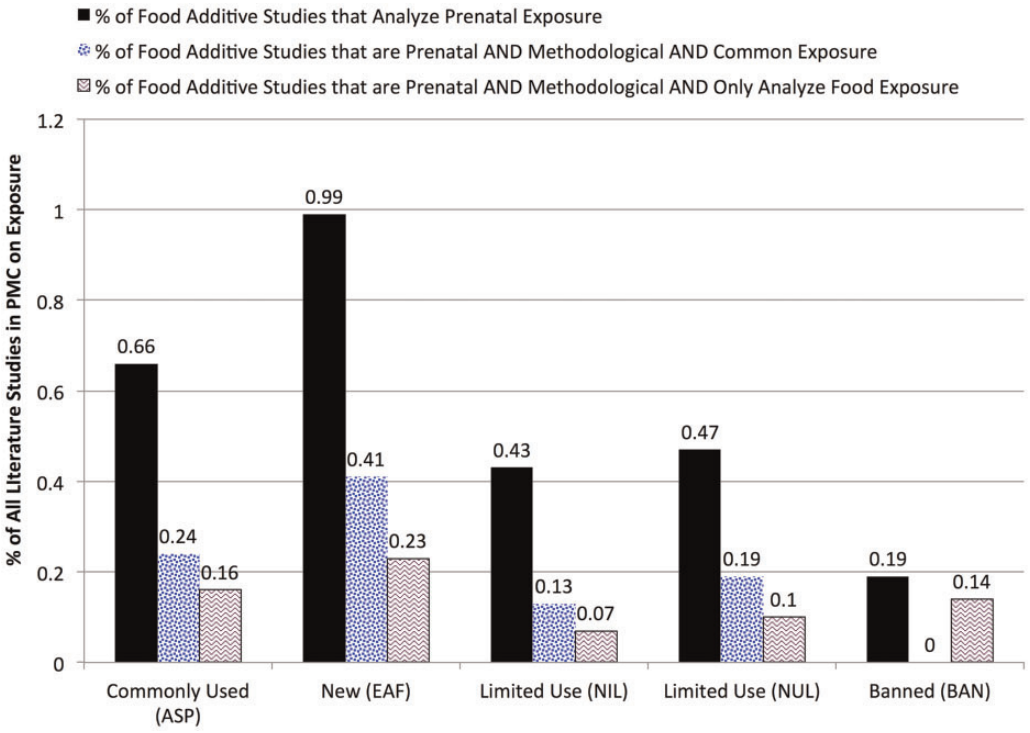

Figure 5. Breakdown of the percentage of all studies in PMC on food additive exposures by study methodology type and food additive type. The percent of food additive studies investigating a prenatal exposure to the additive is provided. We also provide the percent of food additive studies that are methodological and investigate another common exposure (in the 43-exposure set) and also the food additive studies that only investigate a food additive exposure.

Table 3. Breakdown of studies investigating prenatal exposure to food additives by study type

Study Type	All Food Additive Studies (N = 3117)	All Studies (N = 31 764)
Methodology (not including systematic review)	1886 (60.5%)	16 832 (53.0%)
Systematic Review (with methods)	107 (3.4%)	974 (3.1%)
Non-methodology Paper	847 (27.2%)	9060 (28.5%)
PDF only (no method assessment possible)	277 (8.9%)	4898 (15.4%)

investigating the same exposure, namely thiourea (a pesticide). The coumarin study was a PDF-only "state-of-the-art" study from 1985. Thiourea (or ethylene thiourea) is a metabolite of mancozeb. Mancozeb is sprayed on banana plantations in Costa Rica, where pregnant women have been exposed to large doses, because they live in close proximity to the banana plantations.[36] Another study reported elevated thiourea levels in pregnant agricultural rose workers in Ecuador.[37]

Some banned food additives such as calamus root have never been studied (in PMC) in the prenatal exposure context. Calamus root is an herbal remedy used by indigenous peoples of the Americas, especially among the Chipewyan people, to treat fungal infections.[38] It is also noted to have hallucinogenic properties – most likely the reason it was banned as a food additive. However, studying its effects, at least in an animal model setting, is important as individuals may ingest this substance. Without knowledge of the effects of prenatal exposure on the offspring, we would be unlikely to detect this type of poisoning at birth.

PEPPER's main contributions are in identifying literature gaps with regard to methodological studies of prenatal exposure to various compounds, including food additives. In addition, we highlight exposures that are reviewed heavily in the literature with proportionally lower amounts of methodological studies (Table 2). We can also identify exposures where sex-specific outcomes have been investigated heavily vs exposures – including HIV – where very few sex-specific outcomes have been investigated (Figure 3). This form of "deep" analysis of the literature across the breadth of prenatal exposures is important for future researchers, especially given that no comprehensive list of prenatal exposures exists. Our list of 219 exposures can serve as an important starting point for the community while highlighting areas for future research.

There are several limitations of our work. PEPPER provides a list of 219 exposures studied in the literature during the prenatal period. However, PEPPER does not capture every single prenatal exposure studied. PEPPER also depended on manual review of frequent terms. We are aware that this exposure set is not fully complete and requires additional refinement and analysis. Future work includes exploring applications of other methods to learn exposures from PMC.[39] Another limitation is that while PEPPER captures study type (ie methodological, non-methodological, systematic review or PDF-only) it does not capture the outcomes following prenatal exposure. This would require another algorithm to extract the result of the prenatal exposure from the articles and remains a focus of future work.

Table 4. Food additives from EAFUS breakdown by study type

Food Additive	All Studies	Prenatal Exposure Studies	Prenatal Exposure Methodology Studies*	Prenatal Exposure Methodology Studies with Common Exposure AND EAFUS Exposure**	Prenatal Exposure Methodology Studies with Only EAFUS Exposure
Commonly Used (ASP)	463 898	3067	1854	1111	743
New (EAF)	39 821	393	256	163	93
Limited Use (NIL)	5571	24	11	7	4
Limited Use (NUL)	27 493	130	80	52	28
Banned (BAN)	2105	4	3	0	3
Total Unique Studies	445 142	3117	1886	1118	768

*excluding non-methods, systematic reviews.

**lacking other common exposures.

CONCLUSION

In conclusion, we contribute the PEPPER framework that mines full-text articles from PubMed Central and extracts articles investigating exposure to compounds during the prenatal period. PEPPER distinguishes methodology studies from non-methodological studies (eg editorials, perspectives), systematic reviews and PDF-only papers. Overall, only 53.0% of prenatal exposure studies were methodology studies. PEPPER provides a master set of 219 prenatal exposures studied in the literature. No prior comprehensive list existed and therefore PEPPER provides the first such contribution. PEPPER captures exposures from 9492 prenatal exposure methodology studies out of 16 832 studies or 56.4%. PEPPER achieved a recall of 94.4%. PEPPER also highlights important literature gaps, including the paucity of research on prenatal exposure to food additives.

FUNDING

This project was made possible by generous funding from the Perelman School of Medicine, University of Pennsylvania.

COMPETING INTERESTS

None.

CONTRIBUTORS

Conceived Study Design: MRB
Developed methodology: MRB, AK
Performed evaluations: MRB, AK, JX
Assisted with initial review: MRB, JX
Provided informatics and medical advice pertinent to study problem: JH, SL
Wrote Paper: MRB
Reviewed, Edited, and Approved Final Manuscript: MRB, AK, JX, JH, SL

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online. In addition code is available at: https://github.com/bolandlab/PEPPER/

ACKNOWLEDGMENTS

We would like to thank Pam Phojanakong for assistance with the evaluations.

REFERENCES

1. Prüss-Üstün A, Corvalán C. *Preventing Disease through Healthy Environments. Towards an Estimate of the Environmental Burden of Disease*. Geneva: World Health Organization; 2006.
2. Boland MR, Parhi P, Li L. Uncovering exposures responsible for birth season—disease effects: a global study. *J Am Med Inform Assoc* 2018; 25 (3): 275–88.
3. Gardeux V, Berghout J, Achour I, *et al*. A genome-by-environment interaction classifier for precision medicine: personal transcriptome response to rhinovirus identifies children prone to asthma exacerbations. *J Am Med Inform Assoc* 2017; 24 (6): 1116–26.
4. Hanley JP, Jackson E, Morrissey LA, *et al*. Geospatial and temporal analysis of thyroid cancer incidence in a rural population. *Thyroid* 2015; 25 (7): 812–22.
5. Dagliati A, Marinoni A, Cerra C, Gamba P, Bellazzi R. On the correlation between geo-referenced clinical data and remotely sensed air pollution maps. *Stud Health Technol Inform* 2015; 216: 1048.
6. Patel CJ, Bhattacharya J, Butte AJ. An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. *PLoS One* 2010; 5 (5): e10746.
7. Boland MR, Shahn Z, Madigan D, Hripcsak G, Tatonetti NP. Birth month affects lifetime disease risk: a phenome-wide method. *J Am Med Inform Assoc* 2015; 22 (5): 1042–53.
8. Li L, Boland M, Miotto R, Tatonetti NP, Dudley JT. Replicating cardiovascular condition-birth month associations. *Sci Rep* 2016; 6 (1): 33166.
9. Hripcsak G, Knirsch C, Zhou L, Wilcox A, Melton G. Bias associated with mining electronic health records. *J Biomed Discov Collab* 2011; 6: 48–52.
10. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet* 2012; 13 (6): 395–405.
11. Boland MR, Dziuk E, Kraus M, Gelzer A. Cardiovascular disease risk varies by birth month in Canines. *Sci Rep* 2018; 8 (1): doi: 10.1038/s41598-018-25199-w.
12. Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB. Frontiers of biomedical text mining: current progress. *Brief Bioinform* 2007; 8 (5): 358–75.
13. Becker KG, Hosack DA, Dennis G, *et al*. PubMatrix: a tool for multiplex literature mining. *BMC Bioinformatics* 2003; 4 (1): 61.
14. Wei C-H, Harris BR, Li D, *et al*. Accelerating literature curation with text-mining tools: a case study of using PubTator to curate genes in PubMed abstracts. *Database* 2012; 2012: bas041.
15. Frisch M, Klocke B, Haltmeier M, Frech K. LitInspector: literature and signal transduction pathway mining in PubMed abstracts. *Nucleic Acids Res* 2009; 37 (Suppl 2): W135–40.
16. Cheng D, Knox C, Young N, Stothard P, Damaraju S, Wishart DS. PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res* 2008; 36 (Web Server): W399–405.
17. Forbes SA, Bindal N, Bamford S, *et al*. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res* 2011; 39 (Database): D945–50.

18. Szklarczyk D, Franceschini A, Wyder S, *et al*. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 2015; 43 (D1): D447–52.
19. Chen H, Sharp BM. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics* 2004; 5: 147.
20. Zhao D, Weng C. Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction. *J Biomed Inform* 2011; 44 (5): 859–68.
21. Westergaard D, Stærfeldt H-H, Tønsberg C, Jensen LJ, Brunak S. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. *PLoS Comput Biol* 2018; 14 (2): e1005962.
22. Dickersin K. The existence of publication bias and risk factors for its occurrence. *JAMA* 1990; 263 (10): 1385–9.
23. Easterbrook PJ, Gopalan R, Berlin JA, Matthews DR. Publication bias in clinical research. *Lancet* 1991; 337 (8746): 867–72.
24. Vawdrey DK, Hripcsak G. Publication bias in clinical trials of electronic health records. *J Biomed Inform* 2013; 46 (1): 139–41.
25. Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *BMJ* 1997; 315 (7109): 640–5.
26. Jenders RA, Corman R, Dasgupta B. Making the standard more standard: a data and query model for knowledge representation in the Arden syntax. In: *AMIA Annual Symposium Proceedings*. American Medical Informatics Association; 2003: 323.
27. Demner-Fushman D, Hauser S, Thoma G. The role of title, metadata and abstract in identifying clinically relevant journal articles. In: *AMIA Annual Symposium Proceedings*. American Medical Informatics Association; 2005: 191.
28. Winston BH, Sutton RL. Urticaria: detection of ingested, allergens; the single food additive diet. *Practitioner* 1948; 160 (959): 347–52.
29. Helgason T, Jonasson M. Evidence for a food additive as a cause of ketosis-prone diabetes. *Lancet* 1981; 2 (8249): 716–20.
30. Van de Brug F, Luijckx NL, Cnossen H, Houben G. Early signals for emerging food safety risks: From past cases to future identification. *Food Control* 2014; 39: 75–86.
31. Lange S, Probst C, Gmel G, Rehm J, Burd L, Popova S. Global prevalence of fetal alcohol spectrum disorder among children and youth: a systematic review and meta-analysis. *JAMA Pediatr* 2017; 171 (10): 948–56.
32. Wikoff D, Welsh BT, Henderson R, *et al*. Systematic review of the potential adverse effects of caffeine consumption in healthy adults, pregnant women, adolescents, and children. *Food Chem Toxicol* 2017; 109 (Pt 1): 585–648.
33. Mitchell DC, Knight CA, Hockenberry J, Teplansky R, Hartman TJ. Beverage caffeine intakes in the U.S. *Food Chem Toxicol* 2014; 63: 136–42.
34. Fulgoni VL, Keast DR, Lieberman HR. Trends in intake and sources of caffeine in the diets of US adults: 2001–2010. *Am J Clin Nutr* 2015; 101 (5): 1081–7.
35. Plasek JM, Goss FR, Lai KH, *et al*. Food entries in a large allergy data repository. *J Am Med Inform Assoc* 2016; 23 (e1): e79–87.
36. de Joode BW, Mora AM, Córdoba L, *et al*. Aerial Application of Mancozeb and Urinary Ethylene Thiourea (ETU) concentrations among pregnant women in Costa Rica: The Infants' Environmental Health Study (ISA). *Environ Health Perspect* 2014; 122 (12): 1321.
37. Handal AJ, Hund L, Páez M, *et al*. Characterization of pesticide exposure in a sample of pregnant women in Ecuador. *Arch Environ Contam Toxicol* 2016; 70 (4): 627–39.
38. Johnson D, Goward T, Vitt DH. *Plants of the Western Boreal Forest & Aspen Parkland*. Auburn, WA: Lone Pine; 1995.
39. Kilicoglu H, Demner-Fushman D, Rindflesch TC, Wilczynski NL, Haynes RB. Towards automatic recognition of scientifically rigorous clinical research evidence. *J Am Med Inform Assoc* 2009; 16 (1): 25–31.