Exposome Profiling of Environmental Pollutants in Seminal Plasma and Novel Associations with Semen Parameters.

Haotian Wu (hw2694@cumc.columbia.edu)  
Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University

Vrinda Kalia  
Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University

Katherine E. Manz  
School of Engineering, Brown University

Lawrence Chillrud  
Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University

Nathalie Hoffman Dishon  
Infertility and IVF Unit, Department of Obstetrics and Gynecology, Chaim Sheba Medical Center (Tel Hashomer); Sackler Faculty of Medicine

Gabriela L. Jackson  
Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University

Christian K. Dye  
Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University  
https://orcid.org/0000-0003-4532-7042

Raoul Orvieto  
Infertility and IVF Unit, Department of Obstetrics and Gynecology, Chaim Sheba Medical Center (Tel Hashomer); Sackler Faculty of Medicine

Adva Aizer  
Infertility and IVF Unit, Department of Obstetrics and Gynecology, Chaim Sheba Medical Center (Tel Hashomer); Sackler Faculty of Medicine

Hagai Levine  
Braun School of Public Health and Community Medicine, Hadassah Medical Center, The Faculty of Medicine, Hebrew University of Jerusalem

Marianthi-Anna Kioumourtzoglou  
Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University

Kurt D. Pennell  
School of Engineering, Brown University

Andrea A. Baccarelli  
Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University
Ronit Machtinger  
Infertility and IVF Unit, Department of Obstetrics and Gynecology, Chaim Sheba Medical Center (Tel Hashomer); Sackler Faculty of Medicine

Research Article

**Keywords:** exposomics, exposome, seminal plasma, reproduction, reproductive health, semen, sperm, pesticide, etridiazole, NDEA, nitrosamine, 5-ethoxy-3(trichloromethyl)-1,2,4-thiadiazole

**Posted Date:** June 14th, 2023

**DOI:** [https://doi.org/10.21203/rs.3.rs-3058682/v1](https://doi.org/10.21203/rs.3.rs-3058682/v1)

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](https://creativecommons.org/licenses/by/4.0/)
Abstract

There is evidence that indicators of male fertility are in decline globally, but the underlying causes to this pressing global concern have yet to be elucidated. While environmental chemicals are likely major contributors, current knowledge of environmental determinants of male fertility is limited and does not adequately explain this phenomenon. Previous studies have typically examined only limited sets of exposures in blood or urine, which may not accurately capture chemical burden in relevant reproductive tissues, and have overlooked a large range of potential concurrent exposures. Here, using an expansive non-targeted analysis to profile the exposome, we detected widespread environmental pollutant exposure from common and rare environmental exposures in seminal plasma. Machine learning pattern recognition and mixture models identified Etridiazole and N-nitrosodiethylamine to be negatively associated with semen parameters. Our findings suggest that these chemicals are detrimental to sperm and exposomic studies can be used to identify potential reproductive toxicants.

Introduction

Infertility, defined as the inability to achieve a pregnancy after \( \geq 12 \) months of trying, is a global public health concern that affects 15% of all reproductive age couples in the world\(^1,2\). Infertility has significant negative physical, emotional, and social impacts on the affected individuals and addressing infertility will help realize the fundamental right for individuals and couples to have children and to improve the health of the affected individuals. Male infertility contributes to 40–50% of overall cases\(^3\). While changes in gross rates of infertility has not been observed\(^2\), many alarming trends have been observed in male fertility, including notable declines in testosterone and sperm count, with concurrent rises in male genital anomalies and testicular cancer\(^4\). Recent data showed that sperm count declined by 51.6% between 1973 and 2018 and this decline has accelerated since 2000\(^5\). Thus, if this trend persists, a greater proportion of the male population could drop below a threshold\(^6,7\) where lower sperm count could lead to a dramatic rise in male infertility, reflecting not only difficulty in conceiving but also a signal of poorer general health.

The cause of this decline is likely multifactorial, including both prenatal and adult factors\(^8–10\), with environmental factors playing a major role\(^4\). However, our current knowledge of the determinants of male fertility does not sufficiently explain this constant and global decline. Over the past decades, several environmental pollutants have been found to be male reproductive toxicants, such as air pollutants, certain types of persistent organic pollutants, and chemicals commonly found in plastics. However, these known environmental reproductive toxicants might not be the primary drivers of this alarming trend because the overall trends of these specific pollutants are declining across the same timeframe as the observed decline in sperm count and many observed effect sizes are relatively small. Thus, it is likely that there are other unidentified reproductive toxicants or uncharacterized interactions between chemicals.

The emerging field of exposomics offers a much needed comprehensive approach in environmental health research. The concept of the “exposome” was first introduced by C.P. Wild in 2005 as a way to
represent the totality of environmental (i.e. non-genetic) drivers of health and disease\textsuperscript{11}. The exposome is a function of external forces and internal biological processes\textsuperscript{12}. In practice, exposome studies attempt to capture a large set of environmental exposures simultaneously, and have the potential to address limitations of existing studies by investigating the impact of “real life” exposures and their combinations and interactions between exposures.

Historically, there have been several major challenges to identifying previously unknown reproductive toxicants. First, populations around the world are exposed to thousands of chemicals and other environmental exposures with large temporal and geospatial variations, most of which are released without rigorous reproductive toxicity data. This presents an obvious biomonitoring challenge as it is difficult to assess thousands of toxicant exposures and their metabolites simultaneously, particularly chemicals without reproductive toxicity data or those that have never been profiled in humans (“Streetlight Effect”). Second, not every environmental pollutant has accessible assays, creating additional barriers to investigators and biomonitoring efforts. For pollutants that can be measured, the technological platforms used are usually highly specific and lack the ability for high-throughput simultaneous assessment of many exposures from a reasonable amount of biospecimen. This restricts investigators to only a small subset of putative or known toxicants with accessible assays, which impede efforts for non-targeted discovery. This is compounded by a third challenge, i.e., human exposures are typically measured using available and minimally invasive biospecimen such as urine, blood, and hair. This can create exposure misclassification and spurious findings as male reproductive organs and processes are protected by the blood-testis-barrier, which can lead to different exposure profiles between systemic circulation and the male reproductive system\textsuperscript{13,14}. Fourth, the standard statistical approaches make environment-wide (i.e. exposome) association studies (ExWAS) impractical due to multiple testing penalties, leading to high sample size requirements\textsuperscript{15}, similar to genome-wide association studies\textsuperscript{16}. Together, these restraints collectively hindered investigators from more efficiently and comprehensively identifying environmental pollutant influences on male reproductive health.

Our objectives of this study were to 1) profile a large set of organic pollutants relevant to male reproductive tissues using a targeted approach and a discovery-driven non-targeted analysis (NTA) and 2) assess associations between the detected pollutants and semen parameters to discover previously uncharacterized male reproductive environmental toxicants. Our study design, methods, and key findings are shown in \textbf{Fig. 1}. In brief, we leveraged state of the art technological advances in high-resolution mass spectrometry\textsuperscript{17} to simultaneously profile thousands of potential environmental chemicals in seminal plasma, which is more proximal and relevant for male reproductive health compared to measures of chemicals in systemic circulation\textsuperscript{18,19}. We then combined a novel machine learning pattern recognition approach, principal component pursuit (PCP)\textsuperscript{20,21}, with modern statistical mixtures analyses\textsuperscript{22} to efficiently detect associations of environmental chemicals with male reproductive health. Typically, studies model one feature (e.g. genetic loci or environmental exposure) at a time\textsuperscript{23}, and repeated through all features, which incurs severe multiple testing penalties on statistical power. In addition, environmental exposures do not occur in isolation as represented in these models. In real life, we are subjected to
complex sets of environmental exposures that interact with each other to induce endogenous responses. Our new approach helps address both limitations of traditional statistical approaches by removing noise, reducing the number of statistical comparisons, and allowing for discovery of interactions and exposure patterns.

In this study, we found that many organic pollutants and their metabolites are detectable in seminal plasma and show that Etridiazole and N-nitrosodiethylamine (NDEA), both identified with highest level of confidence (level 1), as previously uncharacterized potential human reproductive toxicants. We show that our approach, combining novel machine learning methods and modern statistical models, is consistent with results obtained from traditional statistical approaches, but is more efficient in identifying potential associations using realistic sample sizes.

**Results**

Population characteristics are shown in Table 1. Among the male partners of the 100 couples seeking assisted in vitro fertilization at Sheba Medical Center, five reported use of cannabis and 19 reported current smoking. Their mean age was 38.1 (SD = 5.7) years old and mean BMI was 26.3 (SD = 4.6). The population semen parameter was similar to the WHO population parameters and the majority of participants had normal volume and concentration according to WHO standards. Only 9 couples sought IVF treatment due to male factor infertility.

**Targeted Profiling of the Seminal Plasma Exposome and Exposure Patterns**

Starting with 119 commonly measured persistent or trace organic pollutants (all with Level 1 identification confidence), we found that nearly all tested organic pollutant exposures were detectable in at least one sample, including all measured classes of pesticides, polycyclic aromatic hydrocarbons (PAH), polychlorinated biphenyl (PCBs), and other common environmental contaminants such as dioxins and dioxin-like compounds, phthalates, and solvents (Fig. 2).

Several exposure patterns can be observed in our data (Fig. 2). There were expected correlations between different isomers and related metabolites such as Lindane (also known as γ-hexachlorohexane [HCH]), α-lindane, δ-lindane, and β-HCH. There was also a cluster that comprise numerous pesticides, such as triadimefon, N,N-Diethyl-3-methylbenzamide (DEET), Lindane (and its related isomers and metabolites), 2,4,5-trichlorophenol, Etridiazole, 1,4-dichlorobenzene, and chlorpyrifos, and reagents that are commonly used for production of pesticides, including diethyl phthalate and isosafrole. PCB49 is also part of this cluster of correlated chemicals, which may reflect the usage of PCBs in certain pesticides similar to diethyl phthalate and isosafrole. Other specific co-exposure patterns and exposure profiles were observed among dibenzofuran, safrole, and two PAHs – fluorene and acenaphthene. Although it is not clear which products or behaviors are causing this chemical profile, these four chemicals are also moderately correlated with many pesticides.
We observed predominantly positive correlations, which is consistent with the fact that we expect chemical exposures to arise from complex chemical mixtures and not be mutually exclusive.

**Analysis of Targeted Organic Pollutants Shows Etridiazole was Associated with Poor Semen Parameters**

We used two parallel approaches to quantify the associations of the targeted organic pollutants with semen parameters, including motility, concentration, total motile sperm, and a combined index that captures overall poor semen parameters (Supplemental Fig. 1).

First, we employed an Exposome-Wide Association Study (ExWAS) approach, analogous to Genome Wide Association Studies (GWAS), by testing each exposure individually. Adjusting for age, BMI, smoking and cannabis use, and infertility diagnoses and correcting for multiple comparisons, we found that Etridiazole was associated with lower total motile sperm (FDR q = 0.01), concentration (FDR q = 0.07), and overall index of semen parameters (FDR q = 0.004) (Figs. 3A-B, Supplemental Table 1). These results were robust when we excluded those with male factor infertility or very low total motile sperm count (n = 13), as the association for concentration was below FDR cutoff (q = 0.04) (Supplemental Table 2).

We then sought to use our novel analysis approach, which combines a machine learning pattern recognition approach with modern mixture methods, to identify latent patterns in the exposure data and evaluate these exposure patterns, rather than individual exposures, with the semen parameter outcomes. In this approach, we first used PCP, a robust pattern recognition and data dimension reduction technique, to derive a low-rank exposure matrix that represents the latent underlying patterns without outliers and rare events (Supplemental Fig. 2). Despite the apparent complexity in the observed data, there were ~ 5 principal components (PCP-PCs) that explained > 99% of the variability in the low-rank exposure matrix (Supplemental Fig. 3). Of these 5 PCP-PCs, we found that PCP-PC2 was associated with total motile sperm after adjusting for age, BMI, smoking and cannabis use, and infertility diagnoses and correcting for multiple comparisons (Fig. 3C).

The loadings on PCP-PC2 showed high positive loadings for PAHs and triazine/triazole pesticides and highest negative loadings for Etridiazole and 2,3,7,8-tetrachlorodibenzofuran (Supplemental Fig. 4). To formally investigate which exposures contributed to the association between PCP-PC2 and total motile sperm, the original observed exposure concentrations for the top 10 chemicals with highest loadings were fitted as a mixture in a Bayesian Kernel Machine Regression (BKMR) model adjusting for age, BMI, smoking and cannabis use, and infertility diagnoses. The BKMR model showed that within this mixture, Etridiazole had the highest posterior inclusion probability (Fig. 3D) and was the only one that showed a negative association with any tested semen parameter (Fig. 3E), which is evidence that Etridiazole was the sole driver of this relationship. Indeed, we also saw that Etridiazole was clearly associated with lower sperm motility in the BKMR model, which was also evidenced in the ExWAS models, but did not pass multiple testing corrections.
To ensure that our results were robust and not specific to model parameters, we conducted sensitivity analyses where we specified alternative model parameters for PCP and the number of mixture components for BKMR. There were no appreciable changes in the results as Etridiazole was consistently associated with poor semen parameters (Supplemental Figs. 5–6).

Non-Targeted Analysis Identified NDEA as Associated with Poor Semen Parameters

In addition to the set of targeted organic chemicals, we conducted an NTA of all high-quality spectra tentatively identified in seminal plasma with the goal of applying and leveraging the exposomic approach to identify potential male reproductive toxicants from less investigated pollutants. We characterized the tentatively identified spectra for elemental composition and presence of functional groups (Supplemental Fig. 7). We found that the 39% of the tentatively identified spectra were made of only hydrogen and carbon and ~ 45% did not carry any functional groups. Starting from 814 spectral peaks with abundances detected in all study samples, we again applied our PCP based approach.

In this non-targeted data, PCP derived a low-rank exposure matrix with 6 components that explained > 99% of the variance (Supplemental Fig. 8), one of which was associated with total motile sperm (Fig. 4A). Using BKMR model with the top 10 loading peaks of this component, we found that one peak had the highest posterior inclusion probability for total motile sperm, concentration, and combined index (Fig. 4B). This peak had a retention time of 5.734 minutes and the integrated mass-to-charge (m/z) ratio was 57.03352. Similarly, this peak was the only one in this mixture that showed negative association with semen parameters (Fig. 4C). These results did not differ when we considered alternative BKMR inputs (Supplemental Fig. 9).

To verify the observations from our PCP and BKMR based approach, we also modeled each of the top 10 peaks individually with semen parameters, adjusting for age, BMI, smoking and cannabis use, and infertility diagnoses. These models also show that the same peak (retention time = 5.734 minutes, m/z = 57.03352) was negatively associated with total motile sperm (p = 0.01) and combined index (p = 0.01) and marginally associated with lower sperm concentration (p = 0.07) (Fig. 4D). These results persisted after excluding those with male factor infertility or very low total motile sperm count (p = 0.001 for total motile sperm, p = 0.02 for concentration, p = 0.003 for combined index).

In order to determine the chemical identity of the peak, we used a suspect screening analysis pipeline described previously and identified NDEA as a high quality match for this peak with exact mass match, reverse match factor > 600, and near exact retention time match (within 0.0008%). We compared this peak in our samples to the NDEA reference standard and found very similar chromatograph profile (Supplemental Fig. 10) and ion ratios (Supplemental Table 3). Together, this provides level 1 identification confidence for this peak as NDEA.

Discussion
In this study, we show that numerous organic pollutants can be detected in seminal plasma and can be observed as several pollutants frequently co-occur with a range of other pollutants, forming distinct exposure profiles. Additionally, we applied a novel pattern recognition approach to the exposure data to reduce the dimensionality of the data and subsequently identified Etridiazole and NDEA as two potential male reproductive toxicants negatively associated with semen parameters. These discoveries were supported by the classical ExWAS approach and linear regression analyses, demonstrating the effectiveness and validity of our novel analytical approach, and persisted when men with male factor infertility and low sperm count were removed. On a broader level, our study shows that we can adapt high dimensional approaches to exposomic studies of environmental determinants of reproductive health, which may be necessary to better understand the environmental contributions of the global male fertility decline.

Pesticides have been previously linked to semen parameters. To our knowledge, this is the first study to report the potential association of Etridiazole, a pesticide commonly used to control rot due to fungi and oomycetes, and male fertility parameters. Etridiazole is sold as the active ingredient in numerous commercial pesticides such as Terrazole™, Truban™, and Banrot™ and has been used on golf courses, cotton ornaments, lawns, and agricultural seed products. In addition to occupational exposures, Etridiazole can be present in water and air near sites where it is frequently applied, which may lead to low levels of intermittent exposure in the general population. Etridiazole is not typically found on food, which suggests that the observed exposure is unlikely to be a result of confounding via dietary patterns and associated habits. Etridiazole is classified as a probable human carcinogen, in part due to evidence of testes tumor and testicular interstitial cell hyperplasia in rats, which suggests potential reproductive toxicity. Mechanistically, Etridiazole inhibits fungal growth by hydrolysis of phospholipid membrane of fungal mitochondria via increased phospholipase A activity. The effect of Etridiazole on mammalian cells is not well characterized, but there is evidence that it causes hemolysis and lipid peroxidation of human erythrocyte cell membranes by free radicals. Thus, the biological mechanisms underlying the associations of Etridiazole with semen parameters should be investigated in future studies.

NDEA is a nitrosamine and a probable human carcinogen with well-characterized hepatotoxic, carcinogenic, and mutagenic properties. NDEA is used in a wide range of industrial applications, including usage as gasoline and lubricant additive, antioxidant, plastics stabilizer, and can be found in tobacco and food products. A well-characterized mechanism of carcinogenesis by NDEA is through lipid peroxidation and generation of free radicals. However, lipid peroxidation and free radicals may result in a variety of other adverse health effects, including impaired spermatogenesis and poor semen parameters. Although there was no prior human evidence, there is compelling evidence demonstrating reproductive toxicity of NDEA in experimental models. NDEA increased abnormal sperm and markedly decreased sperm count, sperm motility, and male reproductive organ weight in chronically exposed rats in two different experiments. These changes were complemented by depletion of antioxidants, increase in malondialdehyde and other markers of oxidative stress, elevated indicators of
lipid peroxidation in the testes, pathological changes in seminiferous tubules, and changes in levels of key sex hormones\textsuperscript{39,40}. Similar induction of oxidative stress, changes in sex hormones, and pathological changes in seminiferous tubules were observed in rabbits\textsuperscript{41}. Lycopene, an antioxidant, appeared to rescue the NDEA-induced effects, suggesting that oxidative stress is the underlying mechanism\textsuperscript{39}. Thus, our study is consistent with the emergent experimental evidence and there is strong evidence that NDEA can lead to impaired spermatogenesis and reduced fertility in humans.

In addition to the discovery of two new potential human reproductive toxicants, our study also introduces a new workflow for the analysis of high dimensional exposome data. Our approach leverages a popular machine learning pattern recognition approach to remove outliers and rare events from the noisy exposure data to uncover latent patterns. This allows us to then leverage other machine learning and statistical techniques such as factor analysis (e.g. PCA) and mixtures analysis (e.g. BKMR) to filter hundreds of exposures with minimal number of statistical tests while maintaining the ability to assess potential mixture effects and interactions. As proof of concept, we show that this workflow was able to recapitulate associations that were detectable in classical ExWAS analyses while identifying associations present in our data that were not detectable via classical ExWAS analyses due to lack of power. In other words, our proposed workflow can maximize statistical efficiency and overcome a key challenge in exposome studies\textsuperscript{15} to efficiently screen hundreds of exposures using realistic study sample sizes in reproductive health studies to identify plausible reproductive toxicants.

While our results are promising and consistent with experimental evidence, there are some limitations and areas for future work. First, our study population was recruited from couples seeking IVF in Israel and the associations observed with Etridiazole and NDEA need to be replicated in other populations. However, there is compelling support for the validity of the observed associations from prior literature and experimental evidence. Likewise, while IVF populations are typically not generalizable to the general population, IVF treatment is free to Israeli citizens and many fertile couples take advantage of preimplantation genetic testing, which makes our study population somewhat more similar to the Israeli general population. This is reflected by the relatively few study participants who were diagnosed with male factor infertility and had obvious poor semen parameters. Furthermore, our results were stronger when we excluded individuals with male factor infertility or very low total motile sperm, which gives us confidence that these associations are likely present in the general population as well. Another limitation of our study was that we only assessed semen parameters as the evidence suggested a global decline in sperm count. However, the relationship between semen parameters and fecundity is non-linear\textsuperscript{7,42} and it is unknown whether Etridiazole or NDEA are associated with measures of fecundity such as time-to-pregnancy. Lastly, the mechanism of action, particularly for Etridiazole, is unclear and complementary experimental evidence is necessary to understand how these putative male reproductive toxicants act mechanistically.

Methods
Cohort Recruitment and Clinical Data Collection: One hundred and fifty-eight couples undergoing a fresh IVF cycle with intent to transfer at a Sheba Medical Center, a tertiary medical center in Israel, were approached to join our current study, of which 100 consented. Couples were enrolled during ovarian stimulation and those with severe oligoasthenoteratozoospermia were excluded. Demographic, lifestyle, anthropometric, and medical history data were collected by trained clinic personnel as part of the IVF cycle. Similarly, standard semen parameters, including volume, percent motility, and total sperm count data were collected as part of routine IVF protocol.

Semen samples were collected in a sterile plastic specimen cup after a recommended 2–7 day abstinence period, per standard IVF protocol. Following liquefaction, sperm concentration and motility were evaluated by Makler counting cell (Sefi-Makler). The gradient media was 0.5 ml of 80% PureCeption (Quinn's, ART-2080, SAGE) and 0.5 ml of 40% PureCeption (Quinn's, ART-2040, SAGE). The liquefied semen was gently added to the density gradient medium and centrifuged at 600g for 20 min. The seminal plasma were then aspirated and separated. Samples were stored at –80°C, shipped on dry ice, defrosted in a 4°C refrigerator, and then brought to room temperature prior to extraction for NTA.

High Resolution Non-Targeted Exposome Measurement: Details of the targeted and non-targeted high resolution platform has been described elsewhere\textsuperscript{17}. In brief, seminal plasma was extracted using a modified QuEChERS extraction. A 500 µL volume of seminal plasma was mixed with 500 µL hexane:acetone:dichloromethane, vortex mixed for 30 seconds, and transferred to a 2 mL QuEChERS tube (150 mg MgSO4 + 50 mg C18). The tubes were centrifuged and then the supernatant was transferred to a 2mL glass vial for collection. This was repeated two more time with 250 µL hexane:acetone:dichloromethane. The final extract was reduced to 150 µL under nitrogen, transferred to an amber autosampler vial, spiked with 10 µl of an internal standard solution containing 62.5 µg/L of phenanthrene D-10 and chrysene D-12 (Accustandard) and with 10 µl of diluted retention time marker (Accustandard DRH-TX-003-CNM), and sealed with a cap. The extracts were analyzed using a using a high-resolution Thermo Q Exactive Orbitrap MS equipped with a Thermo Trace 1300 GC and a TriPlus RSH Autosampler. Full details of the analytical methods, including the QC, analytical sequence, chromatography and mass spectrometry settings, and data processing, are provided in the Supplemental Methods. QA/QC provided by the targeted analysis is provided in the Supplemental Methods and Supplemental Tables 4 and 5.

We restricted the analysis of targeted measurements to chemicals with at least 40% detection rate reducing the number of chemicals for analysis from 119 to 43. Of the targeted chemicals profiled, one chemical, Cyanazine, was not detected in any of the samples. Any measurements below the limit of detection were replaced with half the value of limit of quantitation or the minimum detected, whichever was smaller. The non-targeted chemical data analysis was restricted to peaks with an abundance > 50,000 in all samples in order to ensure peak quality, reducing the 2789 peaks detected to 814 peaks used for subsequent analysis.

Chemical Identification and Categorization:
Chemicals measured through the targeted approach were categorized based on entries on the “Chemical Details” provided for each chemical on the EPA CompTox Chemicals Dashboard. They were broadly categorized as: PCB (46), PAH (11), pesticides were sub-classified as chlorobenzene (7), organochlorine (18), triazine and triazole (4), organophosphate (6), and other pesticides (8), and all other chemicals were categorized as “other” (18). To characterize the non-targeted data, we used chemoinformatic tools to generate SMILES by querying the chemical identifier resolver using chemical names, annotated by compound discoverer, using the *webchem* R package (version 1.1.3). We determined the elemental composition and functional groups present using the rcdk toolkit.

For the NTA, peaks were detected with 10 ppm mass tolerance, 10,000 total ion chromatogram threshold, signal to noise ratio of 3, and 99% allowable ion overlap. Each chromatogram was retention time aligned using the carbon distribution marker (contains 9 alkanes; only compounds containing greater than 8 carbons were used since the compounds smaller than this eluted during the solvent delay) spiked into each sample and retention indices (RIs) were calculated for each peak detected.

For suspect screening analysis, the RI of each peak was used to limit suspects during identification; the allowed maximum RI difference was 300. Compounds were identified by searching their mass spectra in the NIST Mass Spectra Library (NIST/EPA/NIH EI and NIST Tandem Mass Spectral Library Version 2.3) and a high-resolution library developed in-house using certified standards containing 354 unique compounds. A minimum Match Factor and Reverse Match Factor score of 500 was used for assigning library matches. Peaks with scores less than 500 were not assigned the identification. Chemicals that matched to our in-house library were assigned Level 1 annotation if they were also detected in our standard mixture, based on previously described scoring scheme.

**Exposome Wide Association Study (ExWAS):** We applied an ExWAS approach to the targeted organic pollutants (exposures with > 40% detection rate) where we tested each chemical individually with the semen parameters as outcomes, analogous to GWAS studies. Using linear regression models, chemicals with 40–60% detection rate were modeled as binary exposures (above and below detection) and those with > 60% detection rate were modeled as continuous variables. This is to balance the additional statistical power gained by modeling variables as continuous when possible and the overall model fit as variables with high non-detects may be less appropriate for linear regression. We also tested different splits (e.g. 40–70% as binary, > 70% as continuous) and it did not change the overall interpretations for the affected variables. Models were adjusted for age, BMI, smoking and cannabis use, and infertility diagnoses (male, female, unexplained). The model estimates were corrected for false discovery rate. As a sensitivity analysis, we excluded 13 individuals who were diagnosed with male factor infertility and/or had total motile sperm count < 10 million.

**Dimension Reduction via Unsupervised Machine Learning Pattern Recognition:** Principal Component Pursuit (PCP) is a robust method for dimensionality reduction and pattern recognition. Its theory and application to environmental health research has been described previously. In brief, we can consider any set of exposure data to comprise of two underlying matrices – the low-rank L-matrix and a sparse S-
matrix. The L-matrix represents the latent underlying patterns that can be estimated, but are not directly observed. The S-matrix represents the unusual or uniquely rare values that cannot be explained by the latent patterns observed by the L-matrix. By partitioning the observed data with PCP, we can effectively separate the underlying latent patterns from the unusual and rare events.

PCP offers two unique advantages. First, by removing the influence of unusual and rare events, we can remove outliers and statistical noise that are often products of random variability. Second, while the resulting L-matrix has the same dimensions as the original exposure matrix, it will have a low rank, effectively reducing the dimensionality of the data. Any subsequent matrix factorization techniques such as principal component analysis (PCA) will identify only a few factors/components that explain a large proportion of variability. This stands in contrast with PCA on typical observed exposure data matrices (i.e. raw exposure data) where a large number of factors/components are required to explain the majority of the variable in the data.

For our presented analysis, we used non-negative non-convex square root PCP with a noise-independent universal choice of regularization parameters. Each non-negative non-convex root square PCP requires two parameters, λ and µ. We tested both default universal parameters \( \lambda = 1/\sqrt{n} \) and \( \mu = p/\sqrt{2} \) as well as parameter values obtained via a grid search. In each case, the resulting L matrices were very similar both in rank and subsequent regression results. Thus, we chose to present the results using the default universal parameters which have been shown to have some theoretical guarantees. Prior to PCP, the exposures were all normalized to 0–1.

PCP was applied on the 43 commonly measured environmental pollutant with > 40% detection rate and 814 universally detected high abundance spectra separately.

Principal Component Regression: We applied PCA on the L-matrices and for downstream regression analyses, we chose the top 5 PCs from the priority common organic pollutants (99.1% cumulative variance) and the top 6 PCs from the non-targeted spectra (99.4% cumulative variance). We extracted the eigenvectors as well as the loadings of each chemical in the dataset. Linear regression was then used to model each principal component as the predictor and semen parameters (total motile sperm, concentration, and percent motility) as outcomes. All models were adjusted for age, BMI, smoking and cannabis use, and infertility diagnoses (male, female, unexplained). Principal components that were found to be significantly associated the outcomes were then chosen for mixtures analysis.

Bayesian Kernel Machine Regression (BKMR): BKMR is a flexible statistical mixtures model that can model multiple exposures simultaneously as a single mixture to estimate the the dose-response relationship of each mixture component while holding other components constant. BKMR makes no assumptions regarding the relationships and can be used to capture non-linear relationships and statistical interactions between the mixture components.

We applied BKMR models to resolve the interpretation issues with principal component regression by taking the top weighted exposures (i.e. greatest absolute loadings) for principal components and treated
them as a statistical mixture in a BKMR model. For all BKMR models, we scaled observed exposures, outcomes, and continuous covariates, and specified 50,000 iterations with default priors. All models were adjusted for age, BMI, smoking and cannabis use, and infertility diagnoses.

We took two approaches to ensure the robustness and validity of our findings. First, we used observed exposure values instead of the decomposed L-matrix. We reasoned that real associations should be detectable using the observed value and that using the latent representations (i.e. L-matrix) may result in spurious findings induced by the data handling procedure and the association is more easily verified and replicated by others. Second, to ensure that our results were not products of arbitrary decisions for the number of mixture components, we tested BKMR models with top 5–15 exposures to ensure that our results were a true reflection of the underlying data.

Linear Regression as Secondary Verification: This analysis was restricted to exposures identified by the PCP-PCA-BKMR pipeline. This is treated as a secondary verification of our PCP-PCA-BKMR results because real associations should be detectable using individual exposure models and potential bias introduced by co-exposure adjustment may lead to spurious findings in statistical mixture models under specific scenarios. Similar to above, these models were adjusted for age, BMI, smoking and cannabis use, and infertility diagnoses.

Declarations

Data Availability: De-identified data from the current study are available from the corresponding author on reasonable request.

Funding: Environment and Health Fund (Israel, RPGA1901 to Levine), National Institute of Environmental Health Sciences (USA, P30ES009089 to AAB, T32ES007272 to KEM),

Acknowledgements: We would like to thank Lukas Scheidl for his assistance with sample extraction.

Author Contributions: HW and VK had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. GLJ was responsible for study logistics and coordination. KEM and KDP were responsible for generating the exposome measurements and provided cheminformatics analyses as well as input on interpretation. LC and MAK reviewed the statistical plan and contributed to the data analysis. CKD contributed to data visualization and interpretation. NHD, RO, AA, HL, and RM were responsible for generating and cleaning the clinical data. HL, RM, and AAB were responsible in acquiring the funding necessary for the study and provided supervision throughout. HW and VK drafted the manuscript. All authors provided intellectual input throughout the project, revised the manuscript, and had final responsibility for the decision to submit for publication.

Competing Financial Interests: The authors declare they have no actual or potential competing financial interests.
References

1. Infertility. https://www.who.int/health-topics/infertility.


50. Wright, J., Ganesh, A., Rao, S., Peng, Y. & Ma, Y. Robust Principal Component Analysis: Exact Recovery of Corrupted Low-Rank Matrices via Convex Optimization. in *Advances in Neural*


Tables

Table 1 is available in the Supplementary Files section.

Supplementary Materials

Supplementary Figures and Tables are not available with this version.

Figures
A. Study cohort

100 couples

Semen parameters measured:
- Total motile sperm
- Motility
- Concentration
- Combined index

B. Exposomic characterization of seminal plasma

Seminal plasma → Cleaning & extraction → GC-based separation → High-resolution mass spectrometry

1. Priority chemicals
   - Targeted analysis
   - 118 chemicals

2. Discovery driven
   - Non-targeted analysis
   - 814 peaks

C. Widespread detection of pollutants in seminal plasma

Priority chemicals

Discovery driven analysis

D. Novel ML approach

PCP based dimensionality reduction

PCA and PC regression to find associations with semen parameters

BKMR to find chemical driving the relationship

Linear regression for secondary verification

E. Discovery of previously unknown associations

**Etradiazole**
- Fungicide that disrupts phospholipid membranes
- Poor data on human health effects

**N-Nitrosodiethy lamine (NDEA)**
- Many industrial applications, uses and sources
- Known to induce lipid peroxidation and oxidative stress

Figure 1

**Overview of the study design, methods, and results.** A. The study population comprise of men from 100 heterosexual couples seeking reproductive health treatment with available semen parameters. B. Using a QuEChERS (quick, easy, cheap, efficient, rugged, and safe) extraction method, targeted and non-targeted organic pollutant exposures were measured from seminal plasma using gas chromatography (GC). C. We detected 118 of 119 organic pollutants in our targeted panel and 814 spectral peaks in non-targeted
discovery analysis. **D.** Novel machine learning pattern recognition methods were combined with modern statistical mixture models and traditional regression results to **E.** identify previously uncharacterized male reproductive environmental toxicants – Etridiazole and N-Nitrosodiethylamine (NDEA).

---

**Figure 2**

High-resolution gas chromatography mass spectrometry efficiently detects widespread detection of organic pollutants in seminal plasma. Left side – the detection rate of 118 targeted organic pollutants. Right – correlation plot of chemicals with ≥40% detection rate.
Figure 3

Exposome-wide association study using standard single-exposure regression models and machine learning pattern recognition based approach both identified Etridiazole as negatively associated with worse semen parameters including lower total motile sperm, percent motility, and concentration. A. Volcano plot of exposome-wide association study results, with the dashed line representing Bonferroni-corrected p-value cut-off threshold. B. Boxplot of semen parameters among those with detectable levels of Etridiazole compared to those with non-detectable levels of Etridiazole. C. Heat plot of results principal component regression. Green indicates negative association and the asterisk denotes statistical significance after Bonferroni correction. D. Bar graph of posterior inclusion probability calculated from Bayesian Kernel Machine Regression (BKMR). E. Estimated dose-response plots and associated 95% credible intervals from BKMR. Green line and credible interval indicates that the credible interval deviates from the null.
Figure 4

**Machine learning pattern recognition based approach identified a spectral peak to be negatively associated with worse semen parameters.** **A.** Heat plot of results principal component regression. Green indicates negative association and the asterisk denotes statistical significance after Bonferroni correction. **B.** Bar graph of posterior inclusion probability calculated from Bayesian Kernel Machine Regression (BKMR). **C.** Estimated dose-response plots and associated 95% credible intervals from BKMR. Green line and credible interval indicates that the credible interval deviates from the null. **D.** Forest plot of estimates and associated 95% confidence intervals from standard single-exposure linear regression models.

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.
• floatimage1.jpeg