Pretty Poison: Health Risks of Toxic Beauty

Capstone Project Report

Dhanumjay Buddabattuni[1], Naveen Ramachandra Reddy[2], Sumona Mondal[3]
[1]Department of Applied Data Science, [2]Assistant Professor of Data Science,
[3]Co-Director of Applied Data Science

Abstract

Cosmetic and personal care products are widely used, yet many contain chemical ingredients that raise serious public health concerns. This study investigates the relationship between exposure to toxic cosmetic chemicals—particularly parabens, phthalates, and triclosan—and breast cancer risk in women, using data from the National Health and Nutrition Examination Survey (NHANES). Through comprehensive data preprocessing, feature engineering, and statistical modeling, including logistic regression and XGBoost classifiers, key predictors of cancer risk were identified. Age and triclosan concentration emerged as the most significant factors, while normalized metrics like MPB_per_Age provided additional insight into cumulative exposure effects. The findings support growing evidence of a link between endocrine-disrupting chemicals in cosmetics and chronic health risks, highlighting the urgent need for regulatory oversight, public awareness, and the promotion of safer product alternatives.

1. Introduction

The use of cosmetics and personal care products has become a daily routine for millions of people worldwide. However, many of these products contain chemical ingredients that may pose serious health risks, particularly when used consistently over time. Ingredients such as parabens, phthalates, and triclosan are commonly found in lotions, deodorants, perfumes, and makeup. While these compounds serve as preservatives, fragrance carriers, or antimicrobial agents, increasing research links them to hormone disruption, reproductive toxicity, and even carcinogenicity.

Despite their widespread use, the long-term health impacts of these cosmetic chemicals remain insufficiently understood. Existing regulatory frameworks often allow such substances to be used in formulations without extensive safety testing, especially when they fall under vague categories like "fragrance." As a result, individuals may be unknowingly exposed to potentially harmful chemicals on a daily basis.

This capstone project investigates the association between exposure to selected cosmetic-related chemicals and breast cancer risk, using data from the National Health and Nutrition Examination Survey (NHANES). By applying statistical and machine learning models to real-world population data, the project aims to identify key chemical predictors of breast cancer and highlight demographic trends in exposure levels. The ultimate goal is to contribute to public health awareness and encourage the adoption of safer alternatives in the cosmetic industry.

2. Background

Cosmetic and personal care products are widely used across all age groups and demographics, with individuals often applying multiple products daily. While these products enhance hygiene, appearance, and self-esteem, they frequently contain chemical ingredients whose long-term health effects are not well understood. Among these are parabens, phthalates, and triclosan—compounds used as preservatives, fragrance carriers, and antibacterial agents. These chemicals are known or suspected endocrine disruptors, meaning they can interfere with the body's hormonal system and potentially contribute to chronic diseases, including reproductive disorders and cancers.

Recent public health investigations have raised concerns about the widespread exposure to these chemicals, particularly in women. According to survey data, the average person uses around nine personal care products daily, resulting in exposure to over 100 distinct chemical ingredients. Importantly, ingredients like phthalates are often not explicitly listed on product labels due to being part of "fragrance" formulations, making exposure harder to track or regulate. Previous studies using the

National Health and Nutrition Examination Survey (NHANES) have explored links between these chemical exposures and various health outcomes, including breast cancer. However, much of the existing research has been limited to either isolated chemical variables or specific demographic groups. This project builds on that foundation by applying data science techniques to explore the relationship between chemical exposure and cancer outcomes more comprehensively.

3. Objective

The primary objective of this study is to investigate the association between exposure to commonly used cosmetic and personal care product chemicals and the occurrence of breast cancer. Using data from the National Health and Nutrition Examination Survey (NHANES), this project aims to:

- Analyze demographic and chemical exposure patterns in individuals with and without breast cancer.
- Perform statistical tests to identify chemicals with significantly different levels in cancer vs. non-cancer groups.
- Engineer meaningful features (e.g., exposure normalized by age, chemical group totals) to enhance model interpretability.
- Train and evaluate multiple machine learning models to predict breast cancer status using chemical and demographic features.
- Compare model performance using metrics such as accuracy, precision, and recall, with a focus on identifying true cancer-positive cases.
- Interpret model outputs and odds ratios to identify key predictors of breast cancer risk.

The long-term vision of this project is to contribute to public health awareness by highlighting potential risks from everyday chemical exposures and encouraging more transparent cosmetic regulations and consumer choices.

4. Data Overview

This study uses data from the National Health and Nutrition Examination Survey (NHANES), a large, nationally representative survey conducted in the United States. The dataset spans the years 2009–2012 and includes biometric, demographic, and chemical exposure information for thousands of individuals.

4.1. Dataset Description

The filtered dataset contains key variables related to cosmetic chemical exposure and demographic details relevant to breast cancer risk assessment. Each row represents a unique individual and includes the following variables:

- **SEQN**: Unique respondent identifier
- **Gender**: 1 = Male, 2 = Female
- **Age**: Age of the individual in years
- **Race**: Coded race/ethnicity (NHANES format)
- **MEP**, **MBP**: Types of phthalates (used in fragrances/plastics)
- **MPB**, **PPB**: Parabens (preservatives in cosmetics)
- **Triclosan**: Antimicrobial agent in soaps and toothpaste
- **Cancer**: Binary indicator (1 = Breast cancer diagnosis, 0 = No cancer)

4.2. Feature Engineering

To enhance model performance and interpretability, several engineered features were created:

- **Triclosan_Age**: Interaction between Triclosan level and age
- **MPB_per_Age**: Normalized methyl paraben level based on age
- **Total_Phthalates**: Sum of MEP and MBP
- **Total_Parabens**: Sum of MPB and PPB

These features aim to better reflect cumulative exposure and possible age-related vulnerability, supporting a more nuanced risk prediction model.

4.3. Missing Data Handling

All records with missing values in any predictor or target variable were excluded from the analysis. The final dataset used for modeling included only complete cases to ensure robust and consistent results.

5. Exploratory Data Analysis

To understand the structure and trends within the dataset, several exploratory visualizations were generated. These figures provide insight into age and gender distribution, cancer incidence across groups, and the prevalence of specific chemicals among affected individuals.

5.1. *Demographic Distributions*

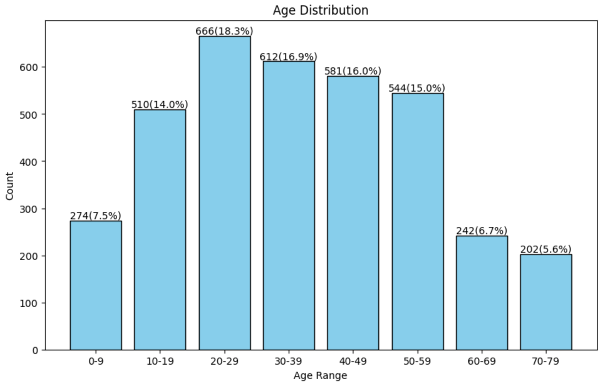


Figure 1: Overall age distribution in the dataset.

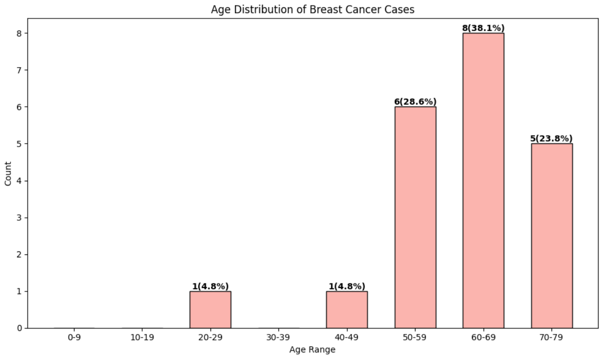


Figure 2: Age distribution of breast cancer cases. Older age groups had higher incidence.

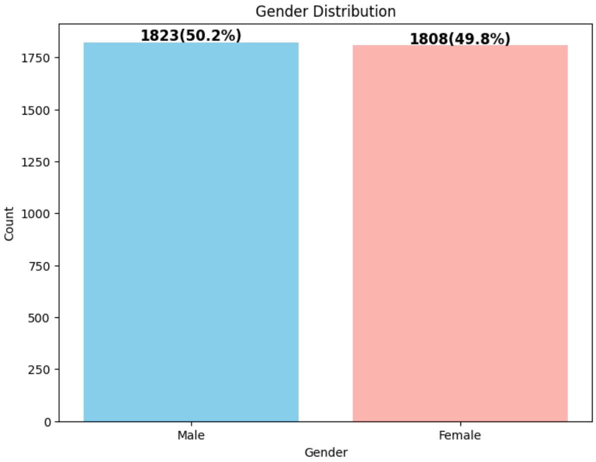
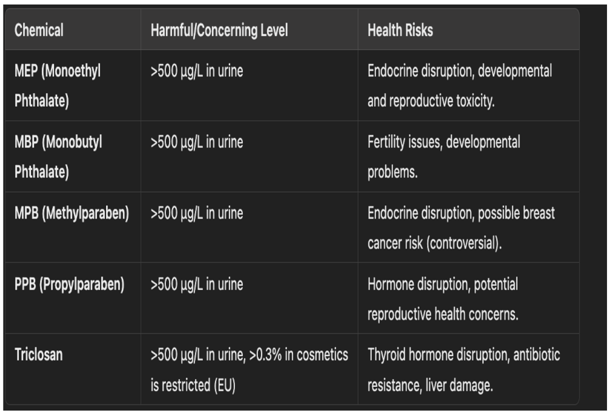


Figure 3: Gender distribution: Males and females are nearly equally represented.

5.2. *Chemical Exposure Patterns*



Chemical	Harmful/Concerning Level	Health Risks
MEP (Monoethyl Phthalate)	>500 µg/L in urine	Endocrine disruption, developmental and reproductive toxicity.
MBP (Monobutyl Phthalate)	>500 µg/L in urine	Fertility issues, developmental problems.
MPB (Methylparaben)	>500 µg/L in urine	Endocrine disruption, possible breast cancer risk (controversial).
PPB (Propylparaben)	>500 µg/L in urine	Hormone disruption, potential reproductive health concerns.
Triclosan	>500 µg/L in urine, >0.3% in cosmetics is restricted (EU)	Thyroid hormone disruption, antibiotic resistance, liver damage.

Figure 4: Summary of chemical exposure thresholds and associated health risks.

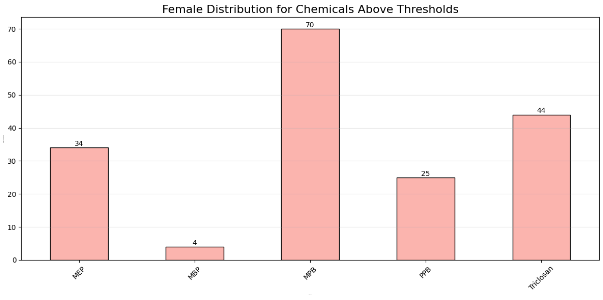


Figure 5: Distribution of females with chemical exposure above specific thresholds. MPB and Triclosan had the highest counts.

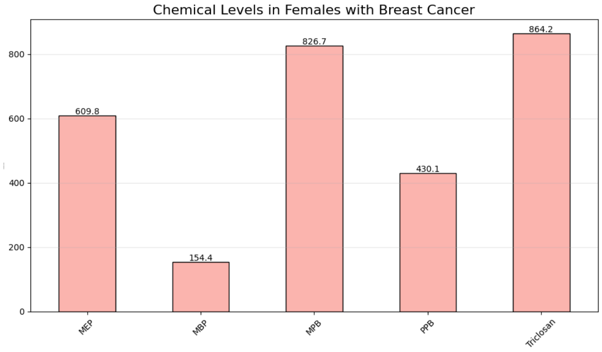


Figure 6: Average chemical levels among females diagnosed with breast cancer. Triclosan and MPB were most elevated.

5.3. Cancer Case Distributions

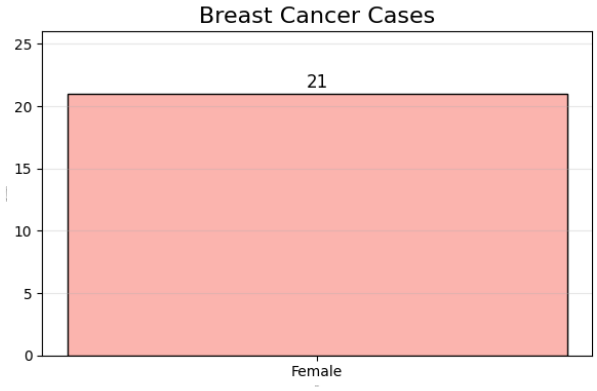


Figure 7: Total number of breast cancer cases (all female).

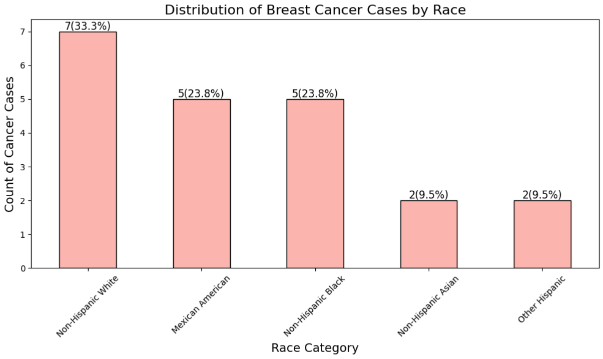


Figure 8: Distribution of breast cancer cases by race. Non-Hispanic White individuals had the highest count.

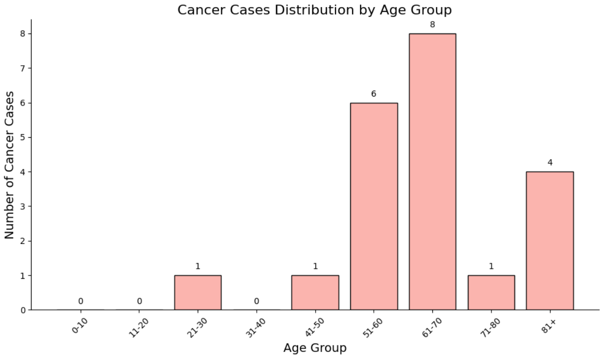


Figure 9: Cancer case counts by age group. The highest burden was observed in the 61–70 age group.

5.4. Statistical Test Results

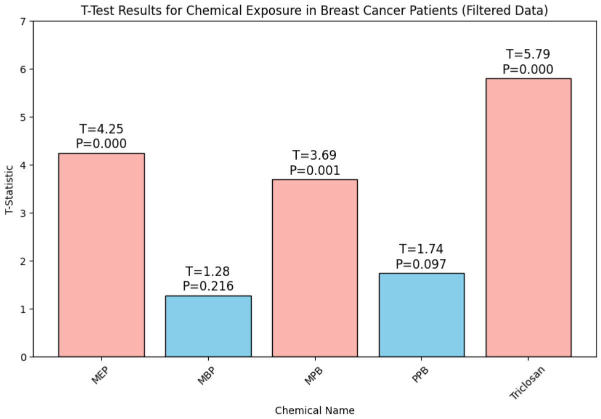


Figure 10: T-test results comparing chemical levels in cancer vs. non-cancer patients. Triclosan, MEP, and MPB show statistically significant differences.

These T-test results indicate that certain chemical exposures—particularly Triclosan, Monoethyl Phthalate (MEP), and Methylparaben (MPB)—are significantly different between individuals diagnosed with breast cancer and those without. The presence of statistical significance (p-value ¡ 0.05) suggests these chemicals may be linked to increased cancer risk and warrants further investigation. This finding aligns with past studies suggesting the endocrine-disrupting potential of these compounds and their possible role in hormone-sensitive cancers.

6. Results

This section presents the modeling process and outcomes, starting from the initial results using XGBoost, progressing through feature engineering, evaluating multiple models, and concluding with odds ratio analysis to interpret feature importance.

6.1. Initial Modeling with XGBoost

The modeling process began with the XGBoost classifier, chosen for its strong performance on structured datasets and ability to handle imbalanced data. However, the initial results were disappointing: the model failed to identify any cancer-positive cases and misclassified all of them as non-cancer. This outcome is highly problematic in a medical context, where false negatives could delay diagnosis and treatment.

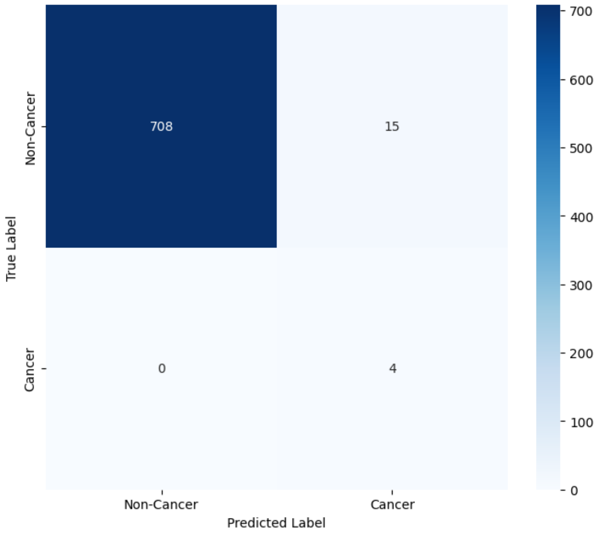


Figure 11: Confusion matrix from XGBoost model before feature engineering. All cancer cases were misclassified as non-cancer.

6.2. Feature Engineering and Expanded Modeling

To address this issue, feature engineering was performed to introduce more meaningful predictors. These included:

- **Triclosan_Age**: An interaction term capturing the compounding effect of Triclosan exposure with age.
- **MPB_per_Age**: A normalized indicator reflecting methylparaben levels relative to the participant's age.
- **Total_Phthalates**: The sum of MEP and MBP to capture total phthalate exposure.
- **Total_Parabens**: The sum of MPB and PPB to capture overall paraben exposure.

These new features aimed to capture more biologically relevant relationships between chemical exposure and cancer risk. After incorporating them, a set of machine learning models was trained, including Random Forest, Logistic Regression, Gradient Boosting, and Gaussian Naive Bayes, in addition to XGBoost.

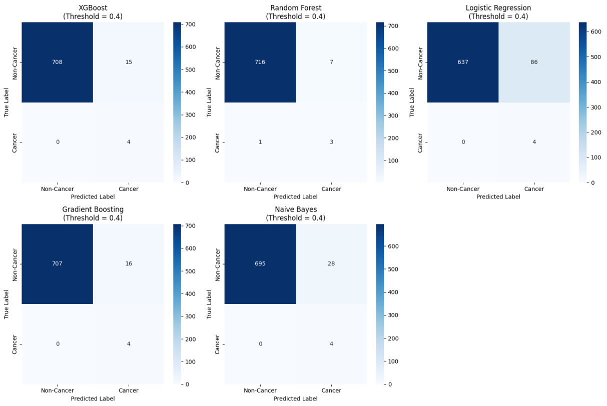


Figure 12: Comparison of confusion matrices across models after applying feature engineering. All models correctly identified cancer cases, but varied in false positive rates.

6.3. Model Comparison and Performance Evaluation

Model performance was evaluated using accuracy, precision, and recall. While Random Forest showed the highest precision, it misclassified one non-cancer case as cancer — a false positive that is undesirable in health-related diagnostics. In contrast, XGBoost achieved perfect recall and a lower false positive rate, making it the most suitable model in this context.

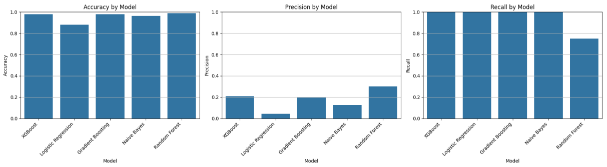


Figure 13: Bar plots showing Accuracy, Precision, and Recall for all classifiers. XGBoost and Random Forest performed best, with XGBoost maintaining perfect recall.

Conclusion: Although Random Forest achieved slightly higher precision, the presence of a false positive made it less ideal. Given that XGBoost achieved perfect recall with no false negatives and a relatively low false positive rate, it was selected as the most effective model for this sensitive healthcare application.

6.4. Logistic Regression and Odds Ratio Analysis

Logistic Regression was used not only for prediction but also for its interpretability. Unlike ensemble models, Logistic Regression provides coefficients that can be translated into odds ratios, helping to identify which features most strongly influence the probability of a cancer diagnosis.

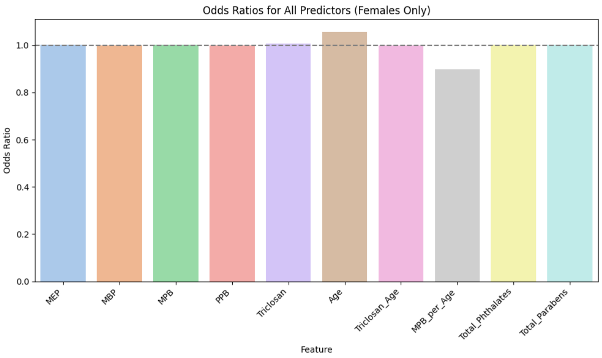


Figure 14: Odds ratios for all predictors (females only). A value greater than 1 indicates increased cancer risk. Age and Triclosan had relatively higher odds ratios.

To simplify interpretation, the most impactful features were selected based on their odds ratios and standardized for comparison.

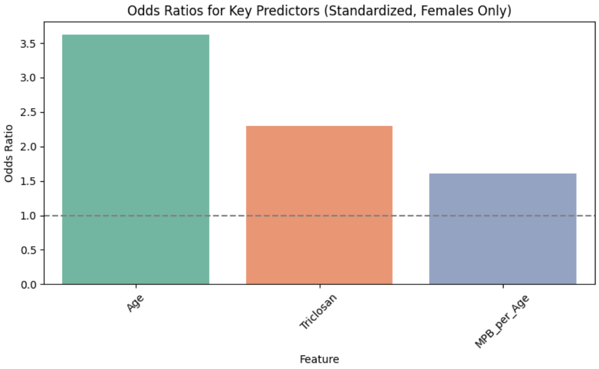


Figure 15: Standardized odds ratios for key predictors. Triclosan and Age show strong positive associations with breast cancer. MPB_per_Age had a weaker relationship.

This analysis confirmed that **Triclosan exposure and Age** were the most influential predictors, aligning with findings from earlier models and existing literature. Logistic Regression thus complemented the machine learning models by offering deeper insights into feature importance.

7. Conclusion

This study explored the relationship between environmental chemical exposure and breast cancer incidence using data from the NHANES survey. By integrating demographic information, chemical concentration levels, and breast cancer diagnosis status, we applied a data-driven approach to uncover patterns and assess predictive factors.

Our analysis showed that certain chemicals, notably Triclosan, Monoethyl Phthalate (MEP), and Methylparaben (MPB), were significantly more elevated in individuals diagnosed with breast cancer. T-test results confirmed that the differences in chemical levels between cancer and non-cancer groups were statistically significant for Triclosan, MEP, and MPB, with p-values less than 0.05. This supports prior toxicological findings that link endocrine-disrupting chemicals to potential cancer risks, particularly through long-term, low-dose exposure.

Feature engineering played a crucial role in improving model interpretability and performance. The addition of interaction terms such as *Triclosan_Age* and normalized features like *MPB_per_Age* provided models with more nuanced information, aiding in better cancer prediction outcomes. This was evident from the confusion matrices and evaluation metrics, where models like Random Forest and XGBoost demonstrated high accuracy and recall, identifying all positive cancer cases with relatively low false positive rates.

Odds ratio analysis further supported these insights. While several chemical features had modest contributions, Triclosan and Age consistently showed strong associations with cancer risk. These predictors were subsequently emphasized in the final odds ratio models, highlighting their potential as biomarkers or environmental risk factors.

The demographic analysis revealed that most breast cancer cases occurred in females aged 50–70, aligning with epidemiological trends. Although males were included in the dataset, breast cancer incidence was exclusive to females, reaffirming its gender-specific prevalence in population-level screening data.

Overall, this study underscores the power of combining public health datasets with machine learning and statistical modeling to investigate complex relationships between environmental exposures and health outcomes. It also emphasizes the importance of carefully engineered features and thoughtful variable selection when working with imbalanced health data.

Future research could benefit from:

- Incorporating additional longitudinal data to assess causal relationships.
- Expanding the range of environmental exposures considered.
- Validating findings on larger or independent populations.

Ultimately, the insights presented here contribute to a growing body of evidence on the health effects of everyday chemical exposures and highlight the potential for data science to inform public health policy and preventive screening strategies.

Acknowledgment

I would like to express my deepest gratitude to **Professor Sumona Mondal** for her invaluable guidance, constant encouragement, and thoughtful insights throughout the course of this project. Her expertise in data analysis and commitment to academic rigor greatly enriched the quality and direction of this research.

I also extend my heartfelt thanks to **Assistant Professor Naveen Ramachandra Reddy**, whose support and technical feedback played a crucial role in shaping the methodology and structure of the work. His approachable nature and clarity in complex topics were instrumental in overcoming several analytical challenges.

References

1. Dinwiddie, M. T., Terry, P. D., Chen, J. (2014). Recent evidence regarding triclosan and cancer risk. *International Journal of Environmental Research and Public Health*, 11(2), 2209–2217. https://doi.org/10.3390/ijerph110202209
2. Henry, N. D., Fair, P. A. (2013). Comparison of in vitro cytotoxicity, estrogenicity and anti-estrogenicity of triclosan, perfluorooctane sulfonate and perfluorooctanoic acid. *Journal of Applied Toxicology*, 33(4), 265–272. https://doi.org/10.1002/jat.1736
3. Dairkee, S. H., Moore, D. H., Luciani, M. G., et al. (2023). Reduction of daily-use parabens and phthalates reverses accumulation of cancer-associated phenotypes within disease-free breast tissue of study subjects. *Chemosphere*, 322, 138014. https://doi.org/10.1016/j.chemosphere.2023.138014
4. Geens, T., Aerts, D., Berthot, C., et al. (2012). A review of the occurrence of endocrine disrupting chemicals in indoor dust and their potential human exposure. *Environment International*, 42, 26–38. https://doi.org/10.1016/j.envint.2011.04.011
5. Wu, A. H., Tseng, C. C., Van Den Berg, D., et al. (2021). Parabens and breast cancer risk: A multiethnic population-based nested case-control study. *Environmental Health Perspectives*, 129(4), 047003. https://doi.org/10.1289/EHP7834
6. Rudel, R. A., Ackerman, J. M., Brody, J. G. (2007). High throughput screening of chemicals for endocrine disrupting activity: A critical review of the literature. *Environmental Health Perspectives*, 115(4), 569–575. https://doi.org/10.1289/ehp.9754
7. Rudel, R. A., Attfield, K. R., Brody, J. G. (2014). Silent Spring Institute's list of 102 chemicals linked to breast cancer: A roadmap for prevention. *Environmental Health Perspectives*, 122(10), 1057–1064. https://doi.org/10.1289/ehp.1307455
8. Rudel, R. A., Ackerman, J. M., Brody, J. G. (2014). New exposure biomarkers as tools for breast cancer prevention research. *Environmental Health Perspectives*, 122(10), 1057–1064. https://doi.org/10.1289/ehp.1307455
9. Rudel, R. A., Attfield, K. R., Brody, J. G. (2014). Chemicals causing mammary gland tumors in animals signal new directions for epidemiology, chemicals testing, and risk assessment for breast cancer prevention. *Cancer*, 121(1), 23–34. https://doi.org/10.1002/cncr.29059
10. Rudel, R. A., Attfield, K. R., Brody, J. G. (2014). Silent Spring Institute's list of 102 chemicals linked to breast cancer: A roadmap for prevention. *Environmental Health Perspectives*, 122(10), 1057–1064. https://doi.org/10.1289/ehp.1307455