# Hazards to Health: The Impact of Areas Added to the National Priorities List (NPL) on Cancer Incidence Rates

MIDN 1/C Nicole Dado and MIDN 1/C Peyton Smith

05 May 2023

SA475E

**Abstract**

The National Priorities List (NPL) was created to address the negative impacts of hazardous waste on the wellbeing of the citizens living in the surrounding area. Using data we obtained from the Environmental Protection Agency and state and local cancer registries, we closely examined whether placement on the NPL signified that a site would have a higher rate of cancer than sites not placed on the NPL. After conducting a difference-in-differences test, we observed that the counties with sites that were on the NPL had higher rates of cancer than counties without sites on the NPL. Future research can expand upon our findings by selecting data that covers a greater span of time, as well as finding regions with hazardous chemical spills that were never placed on the NPL and comparing the cancer incidence rates in those areas to the areas with sites previously on the NPL.

**Keywords:** Difference-in-differences, National Priorities List (NPL), cancer incidence, parallel trends

# 1 Introduction

The National Priorities List (NPL) is a list run by and created by the Environmental Protection Agency. The EPA's goal was to protect U.S. citizens from dangerous chemicals and hazardous waste, specifically after accidents that could put citizens at risk if they were exposed to a high concentration of such waste or chemicals. This goal spurred the creation of the NPL. When a site was added to the list, the EPA supported and helped fund the removal of hazardous substances and pollutants from the site over the course of time, removing a site from the list once the chemicals and pollutants were no longer threatening to the citizens living in the region. The NPL is the most prominent guide used by the EPA in determining which regions need to be investigated further by the agency.

While there exist studies analyzing the impact of hazardous waste on the health of the citizens living near such waste, these studies do not provide evidence of a causal relationship between the presence of waste and cancer incidence. There are also few studies that specifically examine cancer incidence and whether or not an area with a site that has been on the NPL will have a higher rate of cancer. Our goal is to fill in the gaps this research has left behind. The primary question our research aims to answer is whether or not placement on the NPL signifies that a site will have a higher rate of cancer than sites not placed on the NPL. The motivation behind this question is our personal experiences with close family and friends having cancer. We also have a genuine interest in witnessing how the environment we live in impacts our health, especially given that climate change is worsening.

We will examine if there are any significant changes over time in the cancer rate in a given area between the pre-period and post-period, with the post-period being defined by 5 years following a county in the experimental group having a site removed from the NPL (we will dive into what the pre-period and post-period are for the control group later in our paper). This will allow us to have a deeper understanding of the effects of the hazardous waste in the environment that requires a site to be placed on the NPL on the health and well-being of a population. We obtained data from the United States Environmental Protection Agency and local cancer registries on cancer incidence rates and locations of sites on the NPL. We noted which counties had sites in them that were on the NPL at one point, which represents our treatment group, and which counties did not have any sites on the NPL at any point in time, which represents our control group. Using this data, we conducted a difference-in-differences test to determine the effect on cancer incidence of the conditions that cause a site to be placed on the NPL.

We will first discuss literature related to our research, then discuss the empirical strategy we used to answer our research question, followed by an analysis of our results, and finally a discussion of the implications of our results and areas for future research.

## 2 Literature Review

Here, we examine literature relating to our research question, helping inform our decision making and develop a deeper understanding of the topics related to our research.

The Kuznets Curve is a curve that represents the relationship between a nation's wealth and environmental effects. The article being examined, “Environmentally efficient well-being: Is there a Kuznets curve?”, takes this curve further, viewing it from a different angle (Dietz et al. 2012). Dietz’s article utilized the Kuznets Curve to formulate a new model that relates environmental stress and human well being. The results of this work produced multiple curves, the first being an inverted “U” shape and the second being a “U” shape. The inverted “U” shape represents the relationship between inequity and the environment while the “U” represents the relationship between the environment and human well-being. The “U curve” discussion of the Kuznets Curve is extremely applicable to the work that we are doing in addition to the extrapolation of the known theory as this shape indicates that human well being will be high in good environments. We used the results of this paper to predict the results we would obtain in our own research, as environmental changes occur before a site’s removal from the NPL. The authors’ discussed and combatted issues with bias, correlated variables, and the scope (participating areas) in the research and data collection conducted. While we knew other factors could impact cancer incidence, many of these variables would be highly correlated to each other. Our research attempted to combat these issues through the inclusion of fixed effects. We also tried to minimize the number of variables included in our regression. While we handled the correlated variables issue well, we still struggled with data collection.

“The incidence and mortality of esophageal cancer and their relationship to development in Asia” uses statistics on cancer incidence and mortality rates and the Human Development Index to determine if there is a significant relationship between cancer rates in countries in Asia that are either developing or underdeveloped compared to those that are not (Khosravi et al. 2015). The findings of this study indicate that there is a greater incidence of esophageal cancer in countries that are less developed or developing. This is related to our research because it demonstrated that neglected areas or areas that are not receiving the support they need are seeing health implications. These findings are comparable to what we hope our research will uncover, as areas that need to be on the National Priorities List will be impacted by the hazardous chemicals and waste before action can be taken to clean up the area. In our project, we are significantly limited by the number of units of observation we have, which is vastly different from the large amount of data collected and analyzed in the study from the article. Our research also examines cancer incidence through a different lens: the sites impacted by hazardous chemicals, necessitating their placement on the NPL, do receive treatment to address the issues, but these efforts may not necessarily prevent cancer incidence rates from being higher in these areas that have been on the NPL.

In “Cancer Mortality in U.S. Counties with Hazardous Waste Sites and Ground Water Pollution,” the researchers conduct a statistical analysis to determine the relationship between cancer mortality and the presence of hazardous waste (Duncan et al. 1989). These hazardous waste sites are often placed on the NPL due to contamination of groundwater, which is a crucial water supply for the people in these areas. Following chi-squared analyses and the determination of odd ratios, the researchers concluded that more research needed to be done to determine the exact relationship between cancer mortality and sites with hazardous waste. The researchers noted many factors that cannot be controlled for, such as specific lifestyle choices made by the populations studied that affect incidence of cancer, differences in exposure to hazardous chemicals by individuals in the population, and the level of toxicity of chemicals over time. Our research helps to build upon these conclusions, confirming whether or not there are higher rates of cancer in the areas such as those placed on the NPL through a difference-in-differences test. Our empirical strategy forced us to make the assumption that lifestyle choices will be held constant in a given area over the span of time we are researching the area, and these are known as fixed effects. While our study did not focus on mortality rates from cancer in these regions, we look to determine if hazardous waste has an impact on the incidence of cancer in the first place, filling in the gaps left behind by this article.

Along a similar avenue to the above article, the researchers who wrote the journal article “Data Linkage to Explore the Risk of Low Birthweight Associated with Maternal Proximity to Hazardous Waste Sites from the National Priorities List" conducted univariate and multivariate analyses to determine if the proximity of mothers to hazardous waste sites impacted the birthweight of their children while controlling for other variables (Sosniak et al. 1994). The study took a sample from a large population of mothers, computing the unadjusted and adjusted odds ratios for different cases, such as low birth weight and fetal deaths. The results showed that there did not exist a definite relationship between reproductive outcomes for mothers and living close to hazardous waste sites, though there were limitations to the study due to difficulty quantifying the biological exposure of mothers to hazardous waste and differing levels of toxicity and contamination in these hazardous waste sites on the NPL. Our research looked to build upon the findings of this study by determining if the amount of hazardous waste or chemicals that cause a site to be placed on the NPL impacts cancer incidence. This study focused on determining if there was a relationship between childbirth outcomes and a mother’s exposure to hazardous waste and chemicals, which is similar to the analysis we plan to conduct, though our study specifically focuses on cancer. The limitations of this study are similar to those in our research, as we did not have specific data on the level of exposure each citizen had to the chemicals or waste. We did, however, have sites with final site scores that are numerically close to one another when the sites were removed from the NPL, though this did not necessarily mean that each citizen received the same amount of exposure. These site scores quantify the amount of toxicity remaining in the area, and no site is removed from the NPL unless the site score is below a specified threshold. Based on the data, we will be able to state the relationship between our dependent and independent variables assuming the only difference between the groups is being on the NPL or not, with being on the NPL being correlated with having a higher amount of toxic chemicals in the area. In our study, our expectation is that cancer incidence is greater in areas with sites that are added to the NPL. The hazardous waste and its cleanup may not be impacting cancer incidence over time to the extent that we are expecting, though this will be determined from the results of our study. Unlike this article, our project aims to draw a clear conclusion about the impact of hazardous chemicals, which cause a site to be placed on the NPL, on cancer incidence, controlling for many of the factors that this study was unable to control. While our dependent variable is not the same as the dependent variable in this study, our results may lead to a conclusion that could necessitate further research to build upon the findings in this study. Overall, our project aims to draw a clearer conclusion between the variables of interest than this research, which will help to qualify the article further.

While some studies described above were able to find conclusive evidence of a relationship between the independent and dependent variables, the authors of “Hazardous Waste and Health Impact: A Systematic Review of the Scientific Literature” found that this was not always the case (Fazzo et al. 2017). By looking at approximately 60 other research papers, the researchers determined that there was a limited causal relationship being observed between environmental issues and a range of health issues. The causal relationship was deemed not conclusive for the remaining 79 health outcomes being observed. With a closer look, the researchers determined that there were acute effects on a specific population living near an illegal dump site. The researchers also determined which studies to incorporate by rating the hazardous waste and 16 of 95 health outcomes (an assortment of diseases and disorders) on a scale of 1 to 5. Class 5 was the highest, indicating that these studies were very reliable and good to use for this research. This research was extremely impactful on our research and gave us a better understanding of what our outcome may look like. Seeing the inconclusive results was eye opening to the challenges of determining a causal relationship between the environment and cancer rates specifically. This type of research pushed us to slightly tweak our research question in order to compare the cancer incidence of counties with NPL sites compared to those that do not. We took into account some of the flaws from the research analyzed in the paper and built on it in order to look across different groups rather than simply comparing many different sites with hazardous waste.

When determining how long it may take to see the impact of the environment or an environmental change on cancer incidence, we turn to the work of Max Roser in “Smoking: How large of a global problem is it? And how can we make progress against it?” (Roser 2021). In this work of literature, Roser concludes that there is approximately a 20 year lag time between a change and impact on cancer, meaning that it takes 20 years to see the effect of a treatment on cancer incidence. This specific article looks into the lag time between cigarette sales over time and lung cancer mortality. In determining the span of time over which we would collect data, this article was able to provide us insight on how long it would take to see an impact of a change on cancer incidence, informing us in making our decision. While it may be more valuable to increase the time span over which we collected data, we are limited by the availability of data on cancer incidence.

Overall, our research will provide insight on the direct impact of the conditions associated with a site being on the NPL on cancer incidence, expanding upon current literature and generating new conclusions while accounting for the findings in these articles.

# 3 Overview of Data and Empirical Strategy

The observed data in our research comes from 8 counties over the span of 20 years. Of those 8 counties, 5 are in the experimental group and 3 are in the control group. The 3 control groups consist of counties with sites that have never been on the NPL, which allows us to see the rate of cancer incidence without any impact from hazardous waste or chemicals in the surrounding environment. The 3 control counties were Sierra Yuba County, CA, Sandoval County, NM, and Allamakee County, IA. The 5 experimental counties were Lee County, IA, Bergen County, NJ, Oakland County MI, Socorro County, NM, and Kent County, MI, all of which were removed from the NPL in 1995 or 1996. We collected data on cancer incidence in each county 5 years before and 20 years after the site's removal from the NPL for the experimental counties, and we collected the same data for each control county starting in the year 1990 and ending in 2015. All data on cancer was collected from the website Cancer-Rates.info, which provides data from a cohort of states on cancer incidence depending on geographical area, with the option of also grouping by sex and by race/ethnicity. Information from the Kentucky Cancer Registry indicates the legitimacy of our data source, as data from a reliable registry feeds into Cancer-Rates.info ("Kentucky Cancer Registry").

Our data was panel data, meaning we had data on each of the eight sites over the course of 25 years, with data points every five years. The cancer incidence rate, whether or not the site was on the NPL, and whether or not the corresponding entry was from a time before or after 5 years following the removal of the county had a site removed from the NPL are all included in our data file. For the control counties, we chose the years 2000 and after to be considered the post-period because nearly all the counties in the experimental group had the same delineation between pre-period and post-period.

We intentionally chose counties with sites on the NPL that had a minimum time of 15 years between removal and the final year of data collected to account for the lag time associated with a change in cancer incidence that could be correlated with an environmental change. Due to the length of time over which the available data was collected, we were limited to a lag time of 5 years between when a county had a site taken off the NPL and the beginning of our post-period. We were also limited by how far back the available data went. Many counties with sites on the NPL could not be used in our experimental group because the data from their corresponding counties did not cover a large enough period of time.

The research setting for our study is the United States because the 8 counties from which we collected data are located in various states across the country. The results from our research can therefore be applied to all U.S. citizens because we attempted to choose counties from across the U.S., allowing us to draw conclusions to the U.S. population.

As mentioned previously, we chose to conduct a difference in differences test. Difference-in-differences (DID) allows us to compare outcomes between a control group and a treatment group from the same population. With the only difference between the two groups being that the treatment group received the treatment, we can then draw conclusions (assuming parallel trends) on the impact of the treatment on the dependent variable. With DID, in this specific scenario, we cannot make a claim of causality in terms of the impact of the NPL on cancer incidence rates because the counties in our control group do not have hazardous waste or chemicals possibly impacting their cancer incidence. We can, though, draw conclusions about the relationship between the issues that necessitate a site being added to the NPL and the impact they have on cancer incidence in comparison to areas without those issues.

One assumption that needs to be satisfied for a difference-in-difference test is parallel trends. Parallel trends use panel data to show that without a treatment being applied to one group, the control and treatment group would have continued on the same trend in regards to the value of the dependent variable over time. Panel data consists of measurements of multiple units taken over multiple periods of time. In our scenario, the panel data we use provides cancer rates in specific regions over the period of multiple years, with cancer rates being measured multiple times over the course of that specified period. With our panel data, we would assume that the treatment group, which consists of the counties with sites once on the NPL, and the control group, which consists of counties without sites on the NPL, would have differed by the same amount had the treatment not been applied. In this case, the treatment is the presence of hazardous waste on a site and the involvement of the EPA in cleaning the site, necessitating the removal of the site from the NPL. This is an important assumption for our model as it allows us to make conclusions based on the clear differences between the two groups. Without parallel trends, we would not be able to draw these conclusions.

Another crucial aspect of our model is fixed effects. Fixed effects, which are accounted for in a difference-in-differences model, account for variables that we do not directly observe, but have a correlation with the other variables in the model. There are time fixed effects, which are elements that vary over the course of time but not within the individual entities, and unit fixed effects, which are elements that vary within individual entities, but do not vary over the course of time. In our project, an example of time fixed effects amongst the regions would be federal laws and regulations, and an example of unit fixed effects within the regions would be geographic area. By including fixed effects in our model, we are able to control for aspects that may be hard to measure or observe, but remain constant over time or within individual units.

Using the data we collected, we created three variables to be used in our difference-in-differences test: “NPL”, “Post”, and “Cancer.Incidence.Rate”. “NPL” is a binary variable that takes on the value of 1 if a county has a site in it that was on the NPL and 0 otherwise. “Post” is a binary variable that equals 1 if the year is in the post-period, which is after 2000, and 0 otherwise. “Cancer.IncidenceRate” is our dependent variable and represents the percentage of the population with cancer in a specified area. The number of observations, mean, standard deviation, minimum, and maximum for each of the variables is shown below in Table 1.

*Table 1: Summary statistics for variables used in regression analysis*

Data Table

| Statistic | N | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|---|
| Cancer.Incidence.Rate | 32 | 0.610 | 0.136 | 0.363 | 0.984 |
| Post | 32 | 0.750 | 0.440 | 0 | 1 |
| NPL | 32 | 0.625 | 0.492 | 0 | 1 |

As part of our difference-in-differences test, we ran a regression on “NPL”, “Post”, and the interaction between the two variables to predict “CancerIncidenceRate”. The equation accounts for entity and time fixed effects, which are denoted by the two alpha values. The regression equation is shown in Figure 1. The subscript i signifies that the variable varies with respect to county, and the subscript t signifies that the variable varies with respect to time.

$$CancerIncidenceRate_{it} = \beta_0 + \beta_1 OnNPL_i + \beta_2 Post_t + \beta_3 (OnNPL_i * Post_t) + \alpha_t + \alpha_i + \epsilon_{it}$$

*Figure 1: Cancer Incidence Rate Regression Equation*

Using our results, we generated a graph to show the trends in cancer incidence rate for the counties with sites on and not on the NPL from 1990 to 2015. The year 1995 is the year that locations were removed from the NPL, though we indicate years after 2000 as part of the post-period. We chose to put 1995 in the pre-treatment category due to the assumption we added to our model, which was informed via our literature review, that there is a certain amount of time between when a change is made (i.e. being taken off the NPL) and when the impact on cancer incidence occurs at a significant level. Therefore, we would expect the results of the 1995 period to align with that of the pre-treatment period instead of the post-treatment period. We also needed a minimum of two data points to create a trendline, and given that we were significantly limited in the availability data, we had to include 1995 in our pre-period to be able to generate a trendline within that period.

Our model will use the economic principles and techniques mentioned above to determine whether or not cancer rates in a county are higher if the county had a site that was once on the NPL in it.

## 4 Results

The outputs of the regression were obtained using R code. The results from the regression as described are shown in Figure 2.

*Figure 2: Regression summary table results*

| | Dependent variable: |
| --- | --- |
| | Cancer Incidence |
| NPL | 0.002 |
| | (0.098) |
| Post | 0.082 |
| | (0.089) |
| NPL:Post | 0.037 |
| | (0.113) |
| Constant | 0.530*** |
| | (0.077) |
| Observations | 32 |
| R2 | 0.129 |
| Adjusted R2 | 0.036 |
| Residual Std. Error | 0.134 (df = 28) |
| F Statistic | 1.384 (df = 3; 28) |
| Note: | *p<0.1; **p<0.05; ***p<0.01 |

After running the regression on “Cancer.Incidence.Rate” with “NPL” and “Post” as the independent variables, the coefficient on the interaction term, “NPL:Post”, was 0.037, and the coefficient was not statistically significant at the 5% level. This does not mean, though, that there does not exist a relationship between a site being on the NPL and cancer incidence rate. The coefficient is positive, showing that there are higher rates of cancer in counties that have sites that have been on the NPL. The coefficient on “NPL”, while small in size, is also positive and not statistically significant at the 5% level. This coefficient represents the difference between the experimental and control groups before the treatment is applied, which in this case would be having a site within the county that is removed from the NPL. The size and sign of these coefficients signify that in the post-period, a county with a site once on the NPL is predicted to have a higher rate of cancer incidence than a county without a site on the NPL. It is important to note that the sum of the coefficients on “NPL” and “NPL:Post” represent the numerical difference in cancer incidence rates between the experimental and control counties at the end of the post-period.

The graph of our results is shown in Figure 3. The blue trendline, designated by “OnNPL” being “Yes”, corresponds to the counties that once had a site on the NPL and the red trendline, designated by “OnNPL” being “No”, corresponds to the counties without a site on the NPL. The graph has a dotted line on the year 2000 to distinguish between the pre-period and post-period. We observe that after 2000, the trendline for cancer incidence for those areas with sites not on the NPL steadies and drops a bit, then increases again after 2005. The trendline corresponding to the counties with sites once on the NPL is continuously positive and trends upward. The graph shows that even as sites were taken off the NPL, the cancer incidence rates in those counties continued to increase and remain at a higher level than the cancer incidence rates in counties without sites on the NPL. The graph also shows approximate parallel trends in the pre-period, which is an assumption we made that was found to be satisfied by our data.

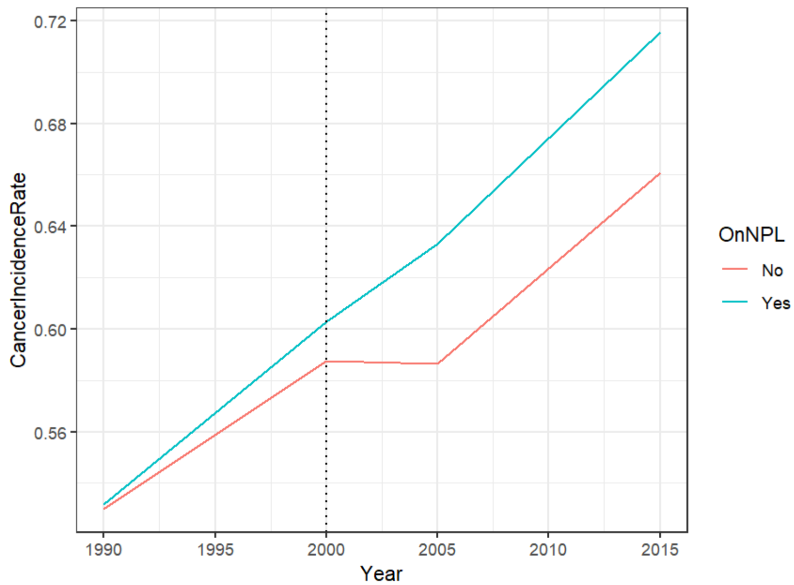


*Figure 3: Graph of cancer incidence rate over time*

We also conducted two fixed effect regressions. The first fixed effects regression was on the year, and the regression results are presented in Figure 4. These year fixed effects account for changes occurring from year to year that are common amongst every county. The coefficients on the three variables that are shown as a factor of a year each quantify the marginal effect of the listed year with respect to 1990. The coefficients on these variables show that the marginal effect of 2015 was the greatest, and the marginal effects of 2000 and 2005 relative to 1990 were similar in magnitude. These effects are accounted for in the regression equation in Figure 1, but running the fixed effects regression is helpful to quantify the changes from year to year that we are unable to observe.

*Figure 4: Summary statistics of regression with year fixed effects*

| | Dependent variable: |
|---|---|
| | Cancer Incidence |
| NPL | 0.030 |
| | (0.048) |
| as.factor(Year)2000 | 0.066 |
| | (0.065) |
| as.factor(Year)2005 | 0.084 |
| | (0.065) |
| as.factor(Year)2015 | 0.164** |
| | (0.065) |
| Constant | 0.513*** |
| | (0.055) |
| Observations | 32 |
| R2 | 0.201 |
| Adjusted R2 | 0.083 |
| Residual Std. Error | 0.131 (df = 27) |
| F Statistic | 1.698 (df = 4; 27) |
| Note: | *p<0.1; **p<0.05; ***p<0.01 |

The second fixed effects regression was on the site, and the regression results are presented in Figure 5. These site fixed effects account for differences between counties that remain constant within each individual county from year to year. The coefficients on the variables that are shown as a factor of a site each quantify the marginal effect of the listed site with respect to Allamakee County, IA. The coefficients on these variables show that the marginal effects of Kent County and Sierra Yuba County were the greatest in magnitude. These marginal effects are accounted for in our original regression equation, and this regression allows us to see and quantify the differences between each county that we may not be able to observe.

*Figure 5: Summary statistics of site fixed effects*

| | Dependent variable: Cancer Incidence |
|---|---|
| Post | 0.105*** |
| | (0.033) |
| as.factor(Site)Bergen County, New Jersey | -0.044 |
| | (0.056) |
| as.factor(Site)Kent County, Michigan | -0.227*** |
| | (0.056) |
| as.factor(Site)Lee County, Iowa | -0.021 |
| | (0.056) |
| as.factor(Site)Oakland County, Michigan | -0.094 |
| | (0.056) |
| as.factor(Site)Sandoval County, NM | -0.019 |
| | (0.056) |
| as.factor(Site)Sierra Yuba County, CA | -0.299*** |
| | (0.056) |
| as.factor(Site)Socorro County, New Mexico | 0.003 |
| | (0.056) |
| Constant | 0.619*** |
| | (0.047) |
| Observations | 32 |
| R2 | 0.747 |
| Adjusted R2 | 0.658 |
| Residual Std. Error | 0.080 (df = 23) |
| F Statistic | 8.470*** (df = 8; 23) |
| Note: | *p<0.1; **p<0.05; ***p<0.01 |

# 5 Discussion

Our results allow us to draw key conclusions, but there are also limitations and areas in which our research could be improved upon in the future. The p-value on each coefficient except the intercept in our original regression equation was fairly high, indicating that the coefficients are not statistically significant. We had a sample size of eight counties, which is quite small and caused us to have a biased estimate for the impact of having a site once on the NPL on cancer incidence. Our small sample size was the main contributing factor to having non-statistically significant coefficients in our difference-in-differences regression equation. With that being said, the high p-values were not indicative of the effect of having a site that was once on the NPL being non-existent, but rather indicated that further analysis needs to be carried out using a larger sample size. Going into this research, we understood that our small sample size could greatly influence our results, which we made sure to note in interpreting our results.

The coefficient on the interaction term from our difference-in-differences regression was positive, indicating that the result we expected did occur. To reiterate, the coefficient on the interaction term was 0.03707, which represents the impact of having a site that was once on the NPL on cancer incidence. This coefficient told us that, based on the data and regression output, cancer incidence changed in counties that had sites once on the NPL within them. Though the coefficient was not statistically significant, the sign and size of the coefficient led to the results we saw in Figure 3, with the experimental and control groups differing in cancer incidence rates. We cannot, however, make a claim about the effectiveness of the NPL in reducing cancer incidence rates. This would have required our control group to be composed of counties with sites that had hazardous waste and chemicals that had yet to be placed on the NPL. With more time, we would be able to investigate further and see if such counties existed, allowing us to determine the causal impact of the NPL itself. We might again run into issues with our sample size, though, but this would only be determined through further analysis.

As previously mentioned, we struggled to obtain large amounts of data and associated data points due to a lack of time, recording, and availability of the data we were interested in collecting. In this paper, we used a 5 year lag from when the site was removed from the NPL and when we began the post-period, with the post-period beginning after 2000 and the final data values coming from the year 2015. With more data in terms of quantity and length of time over which it was collected, we hypothesize that we would see an increased statistical significance for the coefficient on our interaction term. Additionally, this increase in time would allow us to shift the focus on the research towards the effectiveness of the NPL and government policies rather than the environment and its impact on cancer. Increasing the amount of data and length of time over which data was collected would also allow us to consider a greater lag time. It would be interesting to explore how the trend in cancer incidence rates in both the control and experimental groups changes over time using a 15 or 20 year lag, similar to the lag mentioned in the literature review article regarding the effects of smoking.

One specific issue we dealt with was that the concentration of chemicals and hazardous waste across the treatment sites was not consistent. A persistence of chemicals over a longer period of time in one site over another may lead to more lasting hazardous effects from the chemicals even if the EPA has cleaned up both sites to a level that is consistent with their standards. It is important to note, though, that the sites we chose for the experimental group all had similar site scores when they were removed from the NPL, meeting the EPA's threshold. Along a similar line, increased awareness of the impacts of these chemicals and pollutants on cancer incidence may contribute to a rise in the cancer incidence rate, though the impact of this awareness should affect both trendlines in a similar manner.

Compared to the research discussed in our literature review, we looked more closely at the incidence of all cancers rather than a specific type of cancer associated with a specific environmental issue. While we could have decided to analyze the incidence of cancers we knew were more directly related to specific environmental issues, we aimed to draw a broader conclusion, which was partially due to the limitations of our data. Additionally, we used data associated with counties that had sites that were on and were not on the NPL, which other studies did not use.

In conclusion, we see a clear relationship between higher cancer incidence and having environmental issues in a region. A way to improve upon our research would be to conduct the same study after more time has passed, allowing for questions about the effectiveness of the NPL with respect to cancer incidence to be answered rather than whether or not the environment has an impact on medical issues.

# 6 References

Bianchi, Ceccarini, Comba, Della Seta, Fazzo, Martuzzi, and Minichilli. "Hazardous waste and health impact: a systematic review of the scientific literature," https://ehjournal.biomedcentral.com/articles/10.1186/s12940-017-0311-8#citeas, updated 10/11/17, accessed 02/02/23.

Dietz, Thomas, Eugene A. Rosa, and Richard York. "Environmentally Efficient Well-Being: Is there a Kuznets Curve?", https://doi.org/10.1016/j.apgeog.2010.10.011, accessed 02/02/23.

Duncan, Robert, Jack D. Griffith, Wilson B. Riggan, and Alvin C. Pellom. "Cancer Mortality in U.S. Counties with Hazardous Waste Sites and Ground Water Pollution," https://doi.org/10.1080/00039896.1989.9934378, updated 01/01/89, accessed 02/02/23.

"Kentucky Cancer Registry," https://www.kcr.uky.edu/support/cancer-rates/, accessed 02/02/23.

Khosravi, Bahman, Mahdi Mohammadian, Abdollah Mohammadian-Hafshejani, Zohre Momenimovahed, Iraj Pakzad, Reza Pakzad, Hamid Salehiniya, and Shahin Soltani. "The Incidence and Mortality of Esophageal Cancer and Their Relationship to Development in Asia," https://doi.org/10.3978/j.issn.2305-5839.2016.01.11, updated 12/31/15, accessed 02/02/23.

Lee, Christoph and Janie Lee. "Impact of New Technology Adoption on Breast Cancer Screening," https://doi.org/10.1148/radiol.2018182476, updated 03/01/19, accessed 02/02/23.

Roser, Max. "Smoking: How Large of a Global Problem Is It? And How Can We Make Progress against It?", https://ourworldindata.org/smoking-big-problem-in-brief, updated 07/14/21, accessed 02/02/23.

Sosniak, Waldemar, Wendy E. Kaye, and Thomas Gomez. “Data Linkage to Explore the Risk of Low Birthweight Associated with Maternal Proximity to Hazardous Waste Sites from the National Priorities List,” https://doi.org/10.1080/00039896.1994.9937475, updated 07/01/94, accessed 02/02/23.