Analysis of Lead Concentration in Toronto's Drinking Water from 2014 to 2024*

Renfrew Ao-Ieong

January 22, 2024

The consumption of lead is known to have numerous negative effects on human health. Residents of Toronto are able to provide drinking water samples to Toronto Public Health to test for the amount of lead it contains. Using the Non Regulated Lead Sample dataset from the City of Toronto's Open Data Portal, we found that the median amount of lead in Toronto's drinking water has decreased over the past 11 years. This shows good progress towards eliminating lead from our drinking water entirely.

Table of contents

[*] Code and data are available at: https://github.com/RenfrewA/toronto-drinking-water-lead-sample

1 Introduction

Lead is a naturally occurring toxic metal that can be harmful to human health. It can cause neurological, cardiovascular, renal, immunological, reproductive and developmental effects, including developmental neurotoxicity in children. (NCCEH 2022) According to the National Collaborating Centre for Environmental Health (NCCEH), since 1970, Canada has phased out use of lead in gasoline, paint, and other products which has significantly reduced the blood lead levels in Canadians. Currently, the main concern of lead exposure is from drinking water. Presence of lead in drinking water is a serious concern to public health and safety which is why proper precautions are needed to ensure safe drinking water.

Lead can make its way into drinking water when plumbing materials such as pipes, faucets, and fixtures containing lead corrode. In Canada, homes built prior to the 1990s are at risk of using materials containing lead. (Health-Canada 2017) In 2019, Health Canada reduced the Maximum Allowable Concentration (MAC) for lead in drinking water from 10 µg/L to 5 µg/L or 0.01 ppm and 0.005 ppm respectively. (NCCEH 2022) This paper will use parts per million (ppm) as the unit of measuring lead concentration in drinking water.

We examined the dataset containing results from testing water samples obtained by residents of Toronto to see if there was a decrease in the lead concentration in drinking water from 2014 to present. We found that there was a decrease in the median lead concentration in Toronto's drinking water since 2014.

The data section will discuss the process of retrieving, summarizing, and finding meaning from the data. The specifics of how the data was cleaned is included in the Appendix.

2 Data

The data used in this paper was retrieved from The City of Toronto's Open Data Portal through the library `opendatatoronto` (Gelfand 2022). It was then cleaned and analyzed using R, an open source programming language for statistical computing and data visualization (R Core Team 2022) together with the libraries `tidyverse` (Wickham et al. 2019), `dplyr` (Wickham et al. 2023), `janitor` (Firke 2023), `knitr` (Xie 2023), and `ggplot2` (Wickham 2016). The dataset consisted of 12811 individual water samples. Table 1 below shows a sample of the cleaned dataset.

The amount of lead present in water is measured in parts per million (ppm). It is important to note that any higher than 0.005 ppm is not acceptable according to Health Canada. (Health-Canada 2017)

Table 1: Sample of cleaned lead concentration data

Sample ID	Date	Lead Amount (ppm)
1536645	2014-01-01	0.007800
1535456	2014-01-02	0.000110
1536641	2014-01-03	0.000092
1548101	2014-01-06	0.000190
1540991	2014-01-06	0.012000

2.1 Drinking Water Samples

Each sample has a `Sample ID`, `Date` which it was obtained, and `Lead Amount (ppm)` which is the concentration of lead in that water sample in parts per million (ppm). We are interested in the trend of the lead concentration in relation to the year. We have two tables shown below. Table 2 is the mean lead concentration by year while Table 3 is the median lead concentration by year. We obtained the following tables by using the `mutate` function (Wickham et al. 2023) to change the dates to just their year. Then, we used the `summarise` function (Wickham et al. 2023) to obtain 11 rows, each row being a year and the mean or median lead concentration for that year.

Table 2: Mean lead concentration of water samples by year

Year	Lead Amount (ppm)
2014	0.0045282
2015	0.0049639
2016	0.0022060
2017	0.0037887
2018	0.0035105
2019	0.0054675
2020	0.0267115
2021	0.0011348
2022	0.0095108
2023	0.0054486
2024	0.0001450

Table 3: Median lead concentration of water samples by year

Year	Lead Amount (ppm)
2014	0.0003345

Year	Lead Amount (ppm)
2015	0.0003255
2016	0.0001980
2017	0.0001760
2018	0.0001365
2019	0.0001385
2020	0.0000939
2021	0.0001400
2022	0.0001100
2023	0.0002455
2024	0.0001450

2.2 Mean vs. Median Lead Concentration by Year

Plotting the mean and median lead concentration by year using `ggplot2` (Wickham 2016) we can see that there has not been a significant change in the mean concentration of lead in drinking water since 2014. On the other hand there has been a decreasing trend in the median concentration of lead in drinking water since 2014. Figure 1 does not show a significant trend in either direction. This is because of outliers that skew the results. For example, in 2020 the mean concentration was 0.0267115 ppm (More than 5 times the MAC). This was due to a water sample that had a concentration of 8.44 ppm which had a large effect on the mean due to the other values being small numbers such as 0.0005 ppm. Figure 2 which is the median lead concentration by year paints a better overall picture of the water quality in Toronto as the median is not as sensitive to outliers as the mean is.

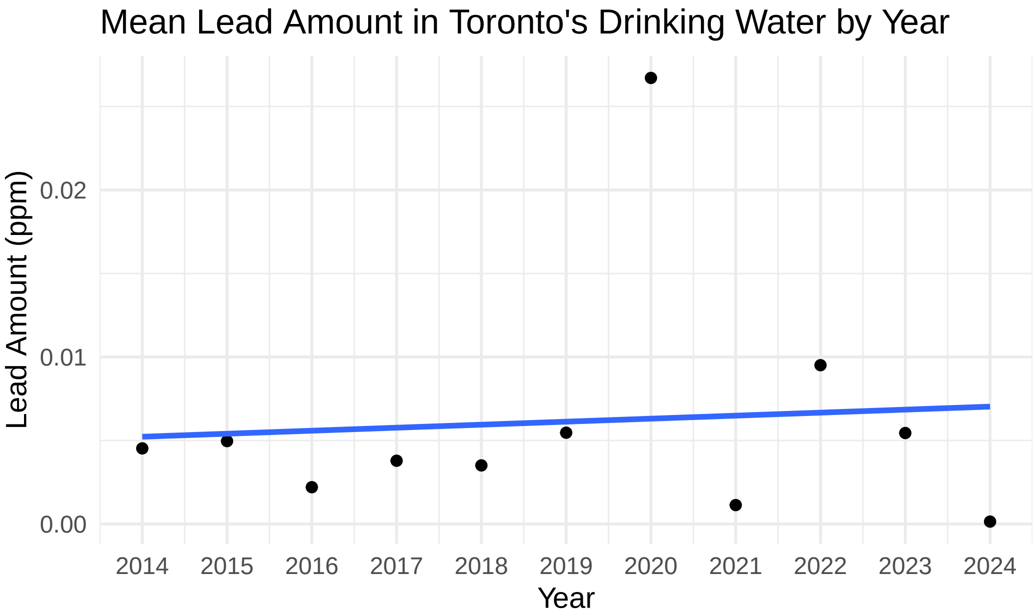

Figure 1: Mean Lead Amount in Toronto's Drinking Water by Year

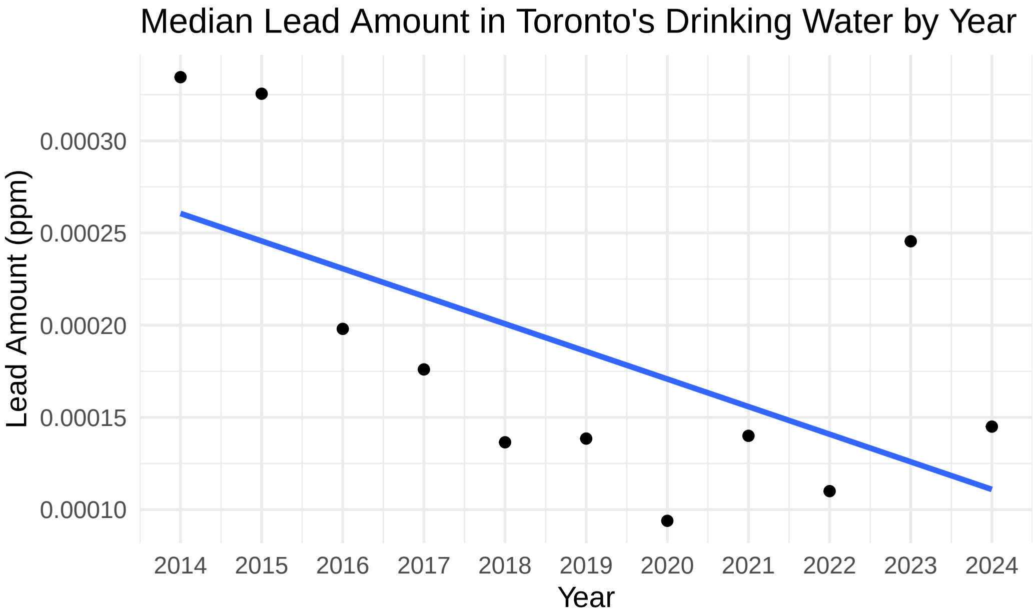

Figure 2: Median Lead Amount in Toronto's Drinking Water by Year

2.3 Limitations

Some limitations to this study is the fact that the dataset contains voluntary samples submitted by residents of Toronto. This may lead to sampling bias as someone who is living in an older home with a higher possibility of containing lead plubmbing materials may be more likely to test their drinking water. Another issue with voluntary samples is that the sample may be tampered with since there are no checks to verify where the water came from. However, it is not feasible for the city to enter people's homes to accurately test their drinking water so this data is the best we have at the moment.

3 References

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Gelfand, Sharla. 2022. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://CRAN.R-project.org/package=opendatatoronto.

Health-Canada. 2017. "Lead in Drinking Water." https://www.canada.ca/en/health-canada/programs/consultation-lead-drinking-water/document.html.

NCCEH. 2022. "Lead in Drinking Water." https://ncceh.ca/resources/subject-guides/lead-drinking-water.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

4 Appendix

4.1 Data Cleaning

The steps taken to obtain this cleaned data is as follows. First, we used the function `clean_names` from the `janitor` package (Firke 2023) to give us column names that are unique and only consist of underscores, numbers, and letters. Next, we converted the `lead_amount_ppm` column from `character` type to `numeric` type. However, there were 2995 samples that contained a lead amount of `<0.00005` which can not be converted directly to `numeric` type as it contains the `<` character. Thus, we chose to convert all samples with that amount to `0.00005`. Then, we chose to discard the `x_id` and `partial_postal_code` columns because those will not be used for our purposes.