Water Environment Research

RESEARCH ARTICLE OPEN ACCESS

A National-Scale Historical Assessment of Nitrate in Public Drinking Water Supplies in New Zealand: Data Integration and Machine Learning Imputation Approaches

Tim Chambers[1] | Frank Dean[1] | Jacques Klavs[2] | Nigel Stanger[2] | Alice Kim[3] | Simon Hales[3] | Jeroen Douwes[4] | Michael G. Baker[3] | Jeremiah Deng[2]

[1]Ngāi Tahu Research Centre, University of Canterbury, Christchurch, New Zealand | [2]School of Computing, University of Otago, Dunedin, New Zealand | [3]Department of Public Health, University of Otago, Wellington, New Zealand | [4]Research Centre for Hauora and Health, Massey University, Wellington, New Zealand

Correspondence: Tim Chambers (tim.chambers@canterbury.ac.nz)

Received: 1 September 2025 | **Revised:** 15 January 2026 | **Accepted:** 20 January 2026

Keywords: drinking water | epidemiology | nitrate | public health

ABSTRACT

Nitrate in drinking water is a known health hazard for infants, although a growing body of epidemiological evidence suggests an increased risk of adverse pregnancy outcomes and some cancers. A major constraint of epidemiological research is the ability to quantify nitrate concentrations in public drinking water supplies over time. Data on nitrate concentrations in public drinking water supplies were retrieved by information requests, linked to a national dataset on the spatial extent of water distribution zones (WDZs) and linked with census information. We applied a number of data cleaning and imputation processes to address complexities in the raw data as well as missingness. In total, 599 WDZs (95.4%) had at least one nitrate measurement between 2000 and 2024 ($n = 20{,}875$ raw observations). After applying a set of imputation methods, the final dataset covered 89.8% of all person-years ($n = 92{,}800{,}000$) of the population on a public drinking water supply during the most recent period from 2000 to 2024. Overall, XGBoost imputation outperformed a range of other imputation methods when synthetic missingness was added to the original data. The large majority (95.3%) of the population was estimated to be on drinking water supplies of less than 1 mg/L nitrate-nitrogen. The population-weighted median nitrate concentration was 0.05 mg/L (IQR 0.04–0.36). This extensive assessment provides the foundation for epidemiological research into the health effects of nitrate contamination of drinking water in New Zealand. The effectiveness of the system for drinking water nitrate surveillance could be enhanced in several ways that would improve its ability to meet its intended purpose.

1 | Introduction

Nitrate (NO_3^-) is a stable form of oxidized nitrogen that occurs naturally and is ubiquitous in the environment (World Health Organization 2016), Nitrate levels in fresh water are affected by agricultural activities (inorganic fertilizers and urine), wastewater treatment, and nitrogenous waste products from humans and discharges from industrial processes (Alam et al. 2024; Fida et al. 2024; Li et al. 2025). The World Health Organization (WHO) guideline value for nitrate in drinking water is 11.3 mg/L nitrate-nitrogen (the unit used throughout this paper) (World Health Organization 2017), which is the equivalent of 50 mg/L nitrate and is based on the absence of adverse health effects caused by infant methemoglobinemia.

A growing body of epidemiological evidence suggests a link between ingested nitrate via drinking water and increased risk of adverse pregnancy outcomes (Royal et al. 2024) and

Summary

- Developed first national-scale nitrate exposure dataset for New Zealand drinking water supplies.
- Coverage achieved for 89.8% of person-years between 2000 and 2024, enabling robust health research.
- Majority of population (95.3%) supplied with water containing less than 1 mg/L nitrate-nitrogen.
- XGBoost imputation method outperformed others for reconstructing missing nitrate data.
- Findings highlight need for consistent monitoring, data standards, and improved nitrate surveillance.

some cancers (Chambers, Douwes, et al. 2022), potentially through the formation of carcinogenic and teratogenic nitrogen-nitroso compounds (NOC) (International Agency for Research on Cancer 2010). In 2020, the French Agency for Food, Environmental and Occupational Health and Safety (ANSES) conducted a systematic review of all case–control and cohort studies investigating associations between nitrate, nitrites or NOC and cancer until March 2022 (ANSES 2022). It concluded that "there is an association between the risk of colorectal cancer and exposure to nitrates and nitrites, whether ingested through drinking water or processed meat"(ANSES 2022, 1) and "in light of new epidemiological and toxicological data, the relevance of the quality limit of 11.3 mg/L nitrate-nitrogen in drinking water"(ANSES 2022, 1) needs to be assessed. In October 2023, the US Environmental Protection Agency (USEPA) announced it would be reviewing nitrate and nitrite toxicity for the first time since its initial assessment in 1991 (United States Environmental Protection Agency 2023).

Whereas the epidemiological evidence of adverse impacts of nitrate in drinking water is growing, few large robust epidemiological studies are available (Chambers, Douwes, et al. 2022; Royal et al. 2024). One constraint has been a lack of longitudinal information on drinking water supplies that can be linked to population health data (Ward et al. 2018). Previous assessments have typically used measured nitrate within a water supply, which is sometimes supplemented by imputation methods (when site values are known) or environmental modeling (when site values are unknown). However, detailed information on the challenges posed by this type of data including the aggregation of data for a water supply with multiple sources or treatment plants as well as decisions on selected imputation methods is often not available.

Approaches to handling missing nitrate data have varied across studies. In a Danish study, around 18% of the final exposure dataset was based on imputed values. Linear interpolation was used when two or more values were available for a site (with no temporal constraint), whereas when a single measurement was available, this was used for the entire period (Schullehner et al. 2017). Researchers in Iowa using the Women's Prospective Cohort used samples collected in periods 1955–1964, 1976–1982, and 1983–1988 to generate an averaged nitrate exposure for the 33-year observation period, effectively imputing missing values within and between periods (Weyer et al. 2001). It is not possible to determine what proportion of the data was imputed, but if there was 100% coverage among the windows above (23 of 33 years), then at least 33% of the annual values are imputed in practice. A Spanish and Italian multicenter case–control study imputed missing values with the average of all available measurements for that water supply, with 12.6% of person-years with a measurement (~90% person-years imputed) (Espejo-Herrera et al. 2016). Lastly, researchers using the Iowa case–control study used a weighted average of adjacent years whereby greater weight was given to years closer in time to the year with no measurements (Ward et al. 2003). This included imputation weights of 1 for 1–2 years from a year with no data, 0.75 for 3–4 years, 0.5 for 5–6 years, and 0.25 for 7–9 years, whereas if there were no data within 10 years, the nitrate level was unknown. A national assessment of nitrate concentrations in drinking water between 2000 and 2020 in France used a combination of hierarchical median imputation (by parameter–year–region) and mixed-effects prediction models ($R^2 = 0.85$) to estimate annual area-level annual values for the approximately 65% missing values (Lafontaine et al. 2024). These more recent assessments demonstrate an advancement in the handling of missing data in epidemiological studies. This study aims to (1) document the methodology of developing a national longitudinal database of nitrate levels in drinking water in NZ and (2) outline the assessment used to quantify nitrate concentrations that will be used in a subsequent epidemiological study aiming to assess associations with adverse pregnancy outcomes.

2 | Methods and Materials

2.1 | Study Design

Longitudinal assessment of nitrate concentrations in public drinking water supplies from 2000 (earliest record) to 2024 in NZ. In NZ, public drinking water supplies are largely managed by 67 territorial authorities (TAs), which are a type of local government responsible for providing local services and regulatory functions within a defined geographic area, including the provision of drinking water.

2.2 | Data

Until 2017, there was no regulatory requirement for a registered supplier to supply their sample results to either Ministry of Health (MoH) or Taumata Arowai—the Water Services Regulator, nor to store and maintain the sample data in a standardized way. In 2017, MoH required water supplies to supply it with their drinking water sample test results. In 2022, Taumata Arowai developed a centralized database (known as Hinekōrako) to which registered suppliers are required to submit their drinking water sample results. This means drinking water quality test samples were only available from a centralized location in a semi-centralized standard since 2017.

Between 2021 and 2024, we sent official information requests to all 67 TAs, the MoH (responsible for drinking water regulation prior to 2021) and Taumata Arowai (responsible for drinking water regulation from 2021). The information requested included (1) any data on the spatial extent of the water distribution

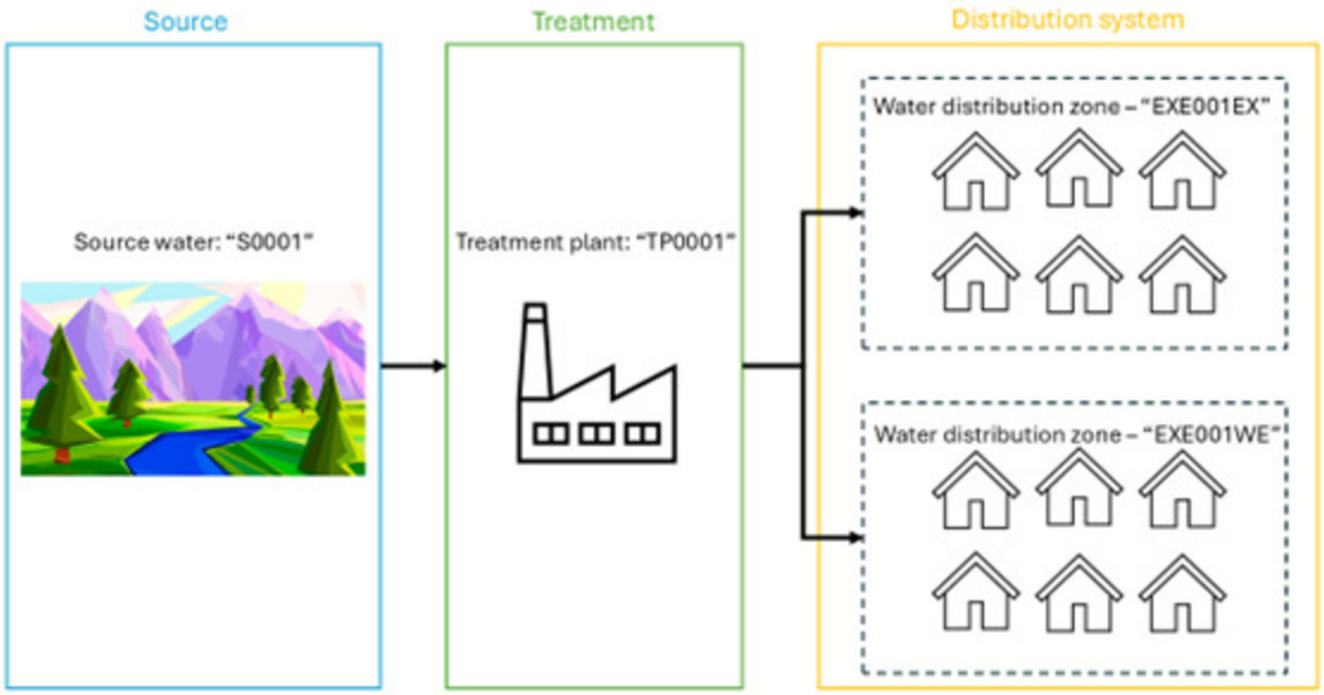

FIGURE 1 | Overview of the relationships between water sources, treatment plants, and water distribution zones with example water supply component codes.

zones (WDZs), the area served by a particular registered water supply; (2) all information about the structure of each of the registered drinking water supplies (e.g., the relationships between water sources, treatment plants and WDZ); and (3) results from all drinking water sampling held on record. Figure 1 highlights the relationships between source waters, treatment plants and WDZ, with example water supply component codes assigned by water regulators in NZ. Importantly, the figure demonstrates that several WDZs can share identical structures (water source and treatment plant—discussed in more detail below).

2.2.1 | Spatial Extent of WDZs

Future epidemiological research will link individual-level residential information to a particular water supply. We have previously collated and standardized the spatial information on 628 TA-owned WDZs (Puente-Sierra et al. 2023). In brief, there was wide variability between TAs in the quality and accuracy of the data provided. We used a combination of data on building footprints and drinking water reticulation (pipe) locations to modify WDZ data provided by TAs. WDZs represent the spatial extent of supplies as of 2022 when the data were collected and do not account for the changing size of these zones over time. As such, using these zones for earlier years may misclassify some people on private supplies as being on public supplies, and vice versa for years after 2022.

2.2.2 | Water Supply Components and Treatment

Taumata Arowai provided information on the supply structure for each of the 628 WDZs for which we had spatial information. This included relationships between sources and treatment plants as well as treatment plants and WDZ. Currently, there is no TA-owned WDZ that uses treatment to remove nitrate.

2.2.3 | Nitrate Data

Drinking water data are sampled at several points within the water supply including at the source (untreated water), at the water treatment plant (after treatment), and at WDZs (in the reticulation system). However, as previously highlighted, coverage for nitrate monitoring in NZ is poor due to there being no requirement for monitoring unless the current levels were above half the MAV, giving a value of 5.65 mg/L nitrate-nitrogen, compounded by the lack of a centralized database until 2017 (Chambers, Hales, et al. 2022).

Extensive detail on the process for cleaning and standardizing these data is provided in the data cleaning protocol (Supporting Information). In brief, we extracted data from various file types using a range of methods. We then identified and standardized all naming conventions for nitrate across the dataset. TAs sometimes test for total oxidized nitrogen as a proxy for nitrate-nitrogen because it combines a total measure of both nitrate-nitrogen and nitrite nitrogen (Land Air Water Aotearoa 2023) and, as incidents of nitrite (and nitrate-nitrogen) are low in New Zealand, the result generally reflects nitrate-nitrogen so we used total oxidized nitrogen as a proxy for nitrate-nitrogen.

2.3 | Generating an Annual Nitrate Value for Each WDZ

For any subsequent health studies, the assessment needs to be aggregated to the WDZ level because residential addresses of

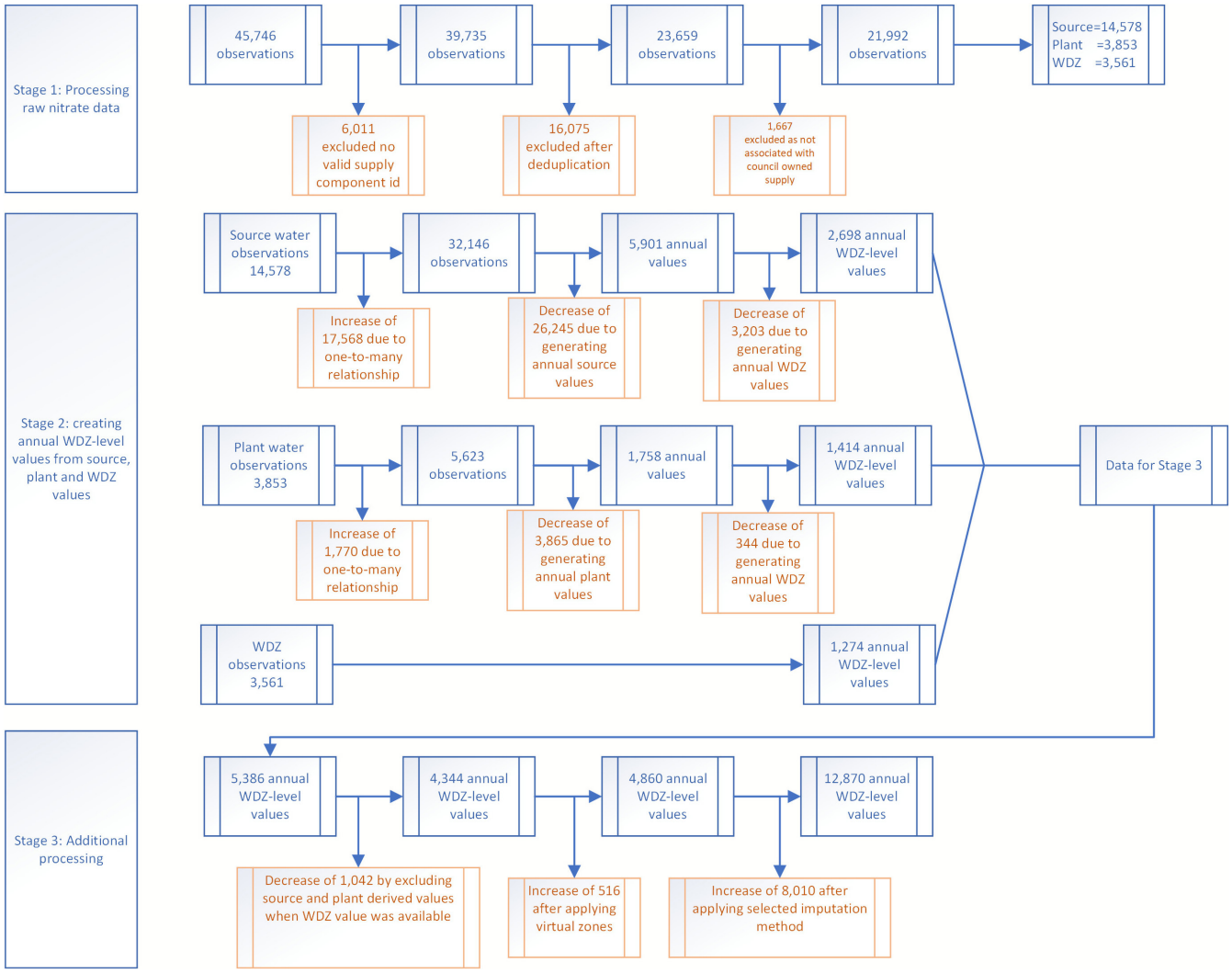

FIGURE 2 | Overview of nitrate assessment data processing methods.

individuals will be spatially linked to a WDZ using the dataset generated by Puente-Sierra et al. (2023). The process to generate an annual value for each WDZ involved three main steps. First, we processed the raw nitrate observations. Second, we generated annual WDZ-level values derived entirely from either source, plant, or WDZ data. Lastly, we ran a series of additional data cleaning processes that involved removing source- and plant-derived values where a WDZ value was available, application of virtual zones, and imputation (Figure 2).

2.3.1 | Stage One: Processing of Raw Nitrate Data

Overall, we received 45,746 nitrate sampling observations from TAs. These were typically grab samples from a water supply component that were processed by an accredited laboratory in NZ. We excluded 6011 observations (13.1% of total raw dataset) because we could not verify or link the reported sample location (e.g., "bore 13") to a Taumata Arowai code (e.g., "S0001"; see Figure 1 for examples). We excluded a further 16,076 observations (35.1%) due to duplicate values by water supply component (e.g., S0001), date (e.g., 01/01/2020), and nitrate value (e.g., 0.1 mg/L nitrate-nitrogen). A further 1667 observations (3.6%) were excluded because they were not associated with one of the current 629 publicly owned WDZs in the spatial dataset compiled by Puente-Sierra et al. (2023). Finally, 1117 observations (2.4%) that were taken prior to 2000 were removed. In total, the processed raw dataset included 20,875 observations (45.6% of the original dataset provided by TAs) with 14,266 source, 3582 plant, and 3027 WDZ observations.

Values below the limit of detection (LOD) formed 4.3% ($n = 903$) of the dataset, with LOD values ranging from 0.002 to 0.5. A substantial majority (94%) of all LOD values were under 0.05. Given the limited clinical meaningfulness of values ranging between 0 and 0.05, based on the current drinking water standard (11.3 mg/L) or epidemiological thresholds for colorectal cancer (> 1 mg/L), we opted for taking the LOD as the measured value.

We conducted an outlier analysis to identify observations that were greater than 2 SD and 1 mg/L nitrate-nitrogen from the mean within sites. Overall, there were only 34 values that meet this definition and contributed < 1% of observations. We decided to include all outliers in the following data processing

steps as they are likely to make a negligible impact on the overall assessment (see Supporting Information for more information).

2.3.2 | Stage Two: Creating Annual WDZ-Level Values From Source, Plant, and WDZ Values

There are complex relationships between source water, treatment plants, and WDZ within drinking water networks. For example, there are one-to-many, many-to-many, and many-to-one relationships between sources, plants, and WDZs. We used the 2006, 2011, and 2024 combined drinking water registries to identify the current and historical relationships between sources, treatment plants, and WDZs over time as well as identify supply components that have changed their official codes over time (for more information, see Supporting Information).

To generate WDZ values from source and plant water values, we followed the same process for source and plant values separately. We first linked raw water values (source, $n = 14{,}578$; plant $n = 3853$) with our relationship tables. Because sources and plants can serve water to multiple WDZ, the total number of observations increased to 32,146 for sources and 5623 for plants. Next, we generated an annual average nitrate level for each unique source and plant within a WDZ so that a WDZ supplied by two or more sources or plants was not disproportionately represented by a component with a higher frequency of testing. Lastly, to create an annual WDZ-level value derived from source or plant water values, we created a simple average of the associated annual source or plant water values for each WDZ. In total, there were 2698 source-derived and 1414 plant-derived annual WDZ-level nitrate values. We had 3561 WDZ nitrate sampling observations that were reduced to 1271 annual WDZ values. In total, there were 5386 annual WDZ values derived from either source ($n = 2698$), plant ($n = 1414$), or WDZ ($n = 1271$). An analysis of the appropriateness of using source and plant observations to infer zone values is provided in Table S2.

2.3.3 | Stage Three: Additional Processing

Three additional processing steps were taken to address some limitations and missing values. First, we excluded all source- and plant-derived annual nitrate values where there was already an observed value for that WDZ. We applied this exclusion because the water in the reticulation is the best representation of the water consumed by the population as it makes the fewest assumptions (e.g., the relative contribution of different sources). This resulted in a reduction of 1042 annual values from 5386 to 4344.

Second, there are 252 WDZs that share the same source and treatment plant as one or more WDZs. This is because they are not actually spatially discrete from another network but represent a pressure zone within a larger network. These 252 WDZs fall within 84 "virtual" WDZs, where the finished water (particularly for non-microbial measures) is likely to be very similar. Consequently, when a value was available for one WDZ within a virtual zone and not another, this value was used for both, resulting in an additional 516 imputed annual values.

2.3.4 | Stage Four: Imputation

Third, in order to impute missing annual nitrate values, we used and compared 13 different imputation methods (median, mean, linear interpolation, polynomial interpolation [2°–4°], spline interpolation [1–4 orders], K-nearest neighbors [k-NN], Bayesian ridge regression, eXtreme gradient boost [XGBoost]). Only the results for six methods (median, mean, spline interpolation [spline: 1], k-NN, Bayesian regression, XGBoost) are displayed in the main text, and others are available in the Supporting Information. Machine learning approaches used auxiliary data on farmed livestock counts and densities, annual total accumulated rainfall and nitrate values from three neighboring zones in the final model. A full description of the methods used to assess different imputation options and results is presented in the Supporting Information.

2.4 | Linkage to Population Served

WDZs have been previously linked to meshblocks (administrative areas with between 60 and 120 residents) for the years 2006, 2013, and 2018 and linked meshblock-level census information on population, ethnicity and neighborhood deprivation (Puente-Sierra et al. 2023). These 628 WDZs provide water to 4.1 million people (based on 2018 census information), which is approximately 85% of the total NZ population and is 99% of people on a registered water supply (Richards et al. 2022). WDZ annual (actual and imputed) concentrations were linked to meshblocks and so can be linked to anyone that received water from a registered water supply operated by TAs in NZ.

Because so few nitrate values existed prior to 2000, we used 2000 as the start of our observation period, continuing through to 2024, to measure data coverage. This provided a total pool of potential person-years of 102.5 million (based on 2018 population census data and 25 years of observation).

2.5 | Statistical Analysis

2.5.1 | Imputation Analysis

Statistical performance metrics and plots used in imputation were generated in Python (Version 3.12). We performed a sensitivity analysis of the imputation methods by evaluating their response to data loss over five increased missingness scenarios (10%, 30%, 50%, 70%, and 90% of original data levels) using synthetic missing completely at random (MCAR) data. We used root mean squared error (RMSE) and R-squared (R^2) for the imputation methods' performance evaluation with RMSE as the primary assessment tool because low variance within-zone means R^2 is sensitive to small prediction errors. We ranked (where the higher the number the better performer) and compared the imputation methods against each other at each level of missingness. We used a density plot to ensure imputed datasets' distribution was true to the original data.

2.5.2 | Population Served Analysis

All descriptive statistics and plots were generated in R (Version 4.2.2). We used linear mixed effects models with a random effect for the WDZ to assess the difference in annual WDZ values in a given year when using the median or mean value as well as to compare annual values derived from source/plant observations compared to those measured at WDZ level. To calculate the population served with a valid nitrate value, we summed the 2018 population by WDZ (Puente-Sierra et al. 2023). To calculate the person-years of observation, we multiplied the total population by the number of years of observation. To estimate the number of people exposed to varying levels of nitrate, we used the most recent sampling result from each WDZ. We categorized exposure into <1.00 mg/L nitrate-nitrogen, 1.00–5.64 mg/L (below half MAV), 5.65–11.2 (above half MAV), and ≥ 11.3 (above the MAV).

3 | Results

3.1 | Observed Nitrate Results Across the Dataset

In total, 599 WDZs had at least one nitrate measurement across the dataset that spanned from 2000 to 2024. The median number of observations across WDZs for the entire period was 7.0 (IQR 2–17), with a minimum of 1 and a maximum of 347 observations. The median number of years of data across WDZs was 5 (IQR 2–10), with a minimum of 1 and a maximum of 23 years. Across the 4000 annual values, the median and minimum number of observations contributing to an annual value was 1.0, whereas the mean count of observations was 2.5 and the maximum was 46 observations.

Figure 3 provides an overview of the population served from those WDZs in any given year, broken down by whether the value was derived from source or plant values, observed at WDZ level, or by virtual zone. Between 2000 and 2012, there was a modest increase in data that covered ~50% of the eligible population. From 2013 to 2023, there was another increase in coverage to around 75% of the population having an annual value. In this period, most values were determined directly through measurements within the WDZ.

3.1.1 | .Imputation Results

Table 1 compares the performance of the selected imputation methods and their ranking relative to the other methods across different missingness ranges. There was no clear standout performer across the imputation methods except mean and median were consistently poorer than the others. Interpolation using spline at order one (spline_1) and k-NN were ranked best for accuracy (RMSE) overall when comparing mean ranking across all levels of missingness (rank = 11.25). Spline_1 was the best performing imputation method at lower levels of missingness (10% and 30%) with RMSE = 0.69 and 0.70, respectively. However, the machine learning methods were more accurate and ranked the best (within 0.13 RMSE of each other) at higher proportions of missingness (50% or more), which is where most of the zones had missing values; 56% of zones have missingness of 50% or more, and of them, 68% of zones have 70% or more missing values.

Spline_1 consistently ranked the best method for explaining variance (R^2) across all missingness ranges although it only performed better at 70% missingness ($R^2 = 0.47$), ahead of XGBoost ($R^2 = 0.44$) (Table S6). However, at lower levels of missingness (10%, 30%, and 50%) spline_1 was only within 0.05 R^2 points of the leading performers, Bayesian regression (at 10%, $R^2 = 0.87$), and k-NN (at 30% and 50%, $R^2 = 0.78$ and 0.52, respectively). XGBoost was at its worst within 0.11 R^2 points of the leaders across the lower missingness levels ($R^2 = 0.82$, 0.67, and 0.45, respectively).

FIGURE 3 | Total population served using 2018 population estimates with a nitrate measurement from 1964 to 2024 by nitrate observation type.

Nitrate-nitrogen imputation reconstruction of the data distribution for all methods imputed datasets (Figure S8) confirms that apart from k-NN, the underlying distribution of the methods compared to the original nitrate-nitrogen dataset remains true with a right skew and a single peak at the lower levels. The k-NN distribution is noticeably different from the other methods with a flatter structure and contains two peaks. This ruled k-NN out of consideration.

XGBoost was the method selected for imputation as it performed better than Bayesian regression and spline$_1$ at the higher levels of missingness and was on par at explaining the variance (R^2).

In total, 8010 annual values were imputed using XGBoost imputation into the final dataset with no temporal threshold for imputation. Overall, the imputed values add an additional 51,685,859 person-years of observation to the exposure dataset. Although imputed values contribute 62% of all annual values, they only contribute 46% of total person-years of observation. Table S7 demonstrates the impact of introducing arbitrary thresholds on the imputation process, for example, introducing a threshold of no more than 5 years between observations would result in the exclusion of 3892 observations, or 38% of the imputed dataset and 30% of the total dataset.

3.1.2 | Final Coverage of Nitrate Assessment and Coverage at Each Step of the Aggregation and Imputation Process

Table 2 outlines the final coverage of the nitrate assessment defined by WDZs (any value), total population (any value), and person-years of observation between 2000 and 2024. When the exposure assessment was restricted to nitrate values measured within the WDZs, there were data for 192 WDZs (30.6%) serving 2,600,000 people (63%) and covering 26,000,000 person-years (25.2%). Coverage of WDZs (94.4%) and population served (99.6%) substantially increased with the inclusion of values derived from nitrate measurements at source or plant, whereas the person-years of observation increased to 39.7%. Inclusion of virtual zone values did not impact the number WDZs or population served but increased person-years by an additional 8.4% of total person-years. Likewise, imputation has a substantial impact on the total person-years of observation, increasing total cover up to 89.8%. In total, 29 WDZs, serving less than 0.4% of the population and covering around 10.2% of all person-years, had no estimates for nitrate during this period.

Figure 4 provides an overview of the person-years of observation by TA from 2000 to 2024 shown in quintiles of coverage. Overall, the map shows that 60% of TAs had less than 25%

TABLE 1 | Imputation method performance measured as root mean squared error (RMSE) and ranking at varying levels and ranges of synthetic missing completely at random (MCAR). Best performer at each level is bolded.

	Percent of synthetic missing at random introduced					Mean rank for different missingness range				
Imputation method	**10%**	**30%**	**50%**	**70%**	**90%**	**All %**	**10%–70%**	**30%–90%**	**50%–90%**	**70%–90%**
Interpolation (spline: 1)	**0.69**	**0.70**	1.36	1.59	4.67	**11.25**	**11.25**	11.00	10.30	10.50
Mean	1.02	0.98	1.58	1.80	4.73	7.60	7.00	8.25	8.00	8.50
Median	1.41	1.17	1.69	2.04	5.00	5.40	5.00	6.00	6.3	6.50
XGBoost	0.83	1.00	1.23	**1.30**	**2.24**	10.40	10.00	11.00	**12.30**	**12.50**
k-NN	0.78	0.82	**1.15**	1.43	**2.24**	**11.25**	11.00	**11.75**	12.00	11.50
Bayesian regression	0.72	1.15	1.26	1.42	**2.04**	10.60	10.00	9.40	12.00	**12.50**

TABLE 2 | The cumulative increase in coverage of the nitrate dataset across the aggregation and imputation methods from 2000 to 2024.

Aggregation and imputation methods	Water distribution zone (any value)		Population served (any value)[a]		Person-years of observation[b]	
Total sample	628	100%	4,100,000	100%	103,400,000	100%
Observed WDZ[c]	192	30.6%	2,600,000	63.0%	26,000,000	25.2%
• Derived	593	94.4%	4,100,000	99.6%	41,000,000	39.7%
• Virtual	599	95.4%	4,100,000	99.6%	49,700,000	48.1%
• Imputed	599	95.4	4,100,000	99.6%	92,800,000	89.8%
No data	29	4.6%	< 100,000	0.4%	10,500,000	10.2%

[a]Rounded to the nearest 100,000.
[b]Calculated from 2000 to 2024 using the 2018 populations associated with WDZ included in this analysis.
[c]Observed WDZ = nitrate value measured at the zone. Derived = a nitrate value derived from source or plant measurements. Virtual = a nitrate based on zone value from a zone with an identical supply structure (same sources and plants). Imputed = a nitrate value imputed based on XGBoost imputation.

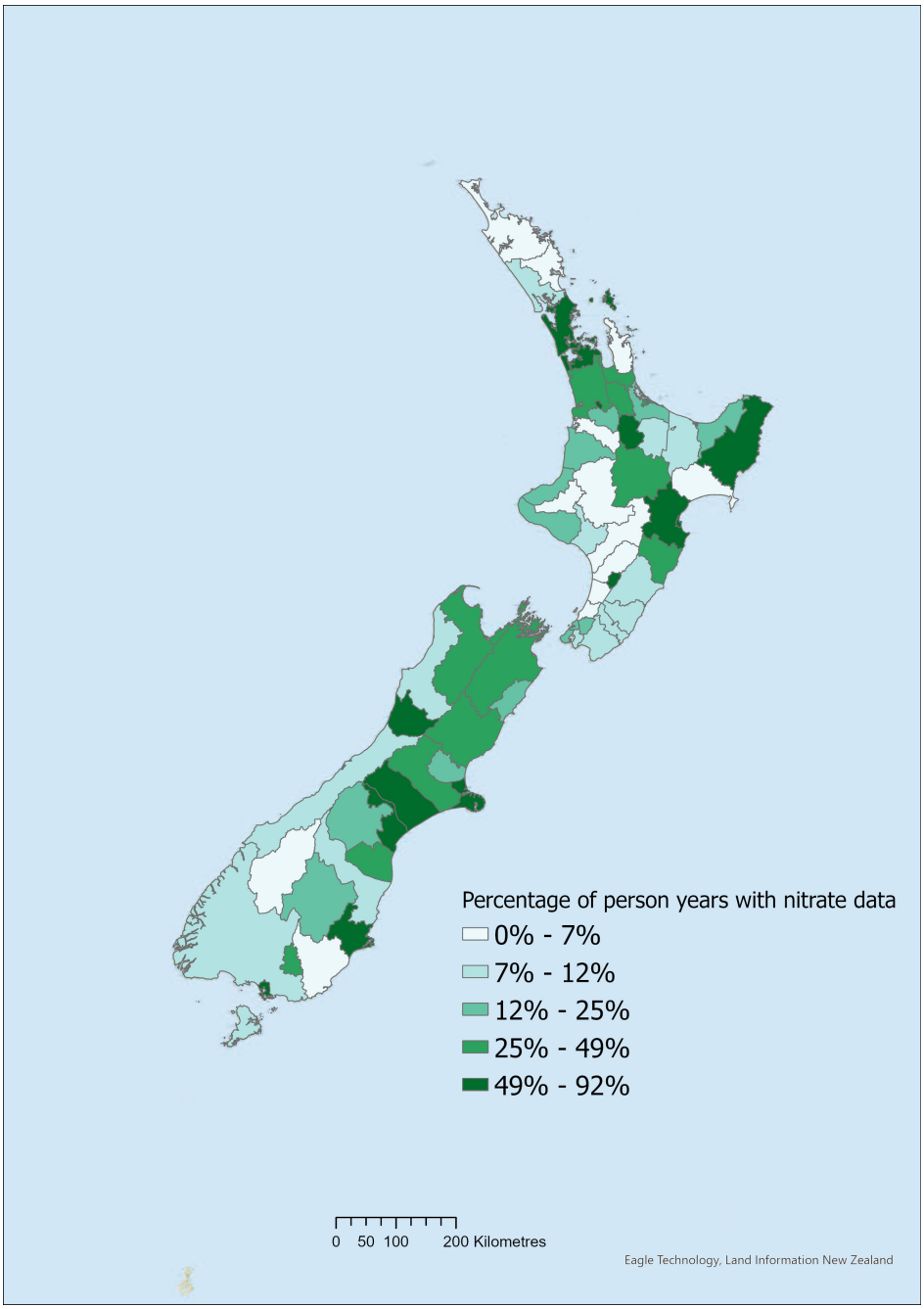

FIGURE 4 | Percentage of person-years with a nitrate measurement between 2000 and 2024 across territorial authorities in New Zealand (see observed WDZ values in Table 2).

coverage in terms of person-years across this period. A further 20% had coverage of between 25% and 49% and the top quintile was between 49% and 92%. In general, larger urban TAs such as Auckland, Christchurch, and Invercargill had greater coverage than smaller rural TAs.

3.1.3 | Distribution of Nitrate Exposure on Registered Supplies

Table 3 provides an overview of the population exposed to varying levels of nitrate concentrations in registered council-owned WDZ using the final dataset including imputed values. The values represent the most recent sample collected from each WDZ. In total, 81.2% of WDZs serving 95.3% of the population have a nitrate concentration of less than 1 mg/L nitrate-nitrogen. An additional 13.1% of WDZs serving 4.2% of the population have nitrate concentrations above 1 mg/L and below 5.65 mg/L (half the MAV). But an additional 1.1% of WDZs serving water to < 1% of the population have nitrate concentrations above 5.65 mg/L (including those above the MAV). The median nitrate concentration among WDZ was 0.16 (IQR 0.04–0.60) and mean 0.69 (SD 1.99). The population-weighted median was 0.05 (IQR 0.04–0.36), and mean was 0.30 (SD 0.72), which were lower than WDZ-based summary statistics.

TABLE 3 | Distribution of nitrate concentrations across water distribution zones owned by territorial authorities and the population served by those zones in the final dataset.

mg/L nitrate-nitrogen	WDZ	%	Population served	%
Total	628	100	4,135,000	100
< 1.00	510	81.2	3,941,000	95.3
1.00–5.64	82	13.1	173,000	4.2
5.65–11.2	4	0.6	2000	< 0.1
> = 11.3	3	0.5	1000	< 0.1
No data	29	4.6	18,000	0.4
Summary statistic	WDZ-weighted concentration (mg/L nitrate-nitrogen)		Population-weighted concentration (mg/L nitrate-nitrogen)	
Mean (SD)	0.69 (1.99)		0.30 (0.72)	
Median (IQR)	0.16 (0.04–0.60)		0.05 (0.04–0.36)	

Abbreviation: WDZ, water distribution zone.

4 | Discussion

Our study has detailed the development of a national assessment tool for nitrate contamination in publicly owned drinking water supplies. After applying several processing steps, 11,289 annual values for WDZs covering 111 million person-years of observation were available for the 1964–2024 period. This includes imputed values that contributed 46.3% of the entire dataset. From 2000 onwards, almost all publicly owned WDZs serving 99.6% of the population on a registered supply have at least one nitrate measurement, whereas estimates included 89.2% of person-years. Overall, the large majority of WDZs (81.2%) and people (95.3%) have nitrate concentrations below 1 mg/L nitrate-nitrogen in their publicly owned drinking water supplies.

Imputation methods are common among similar nitrate assessments in epidemiological studies (Espejo-Herrera et al. 2016; Schullehner et al. 2017; Ward et al. 2003; Weyer et al. 2001). However, it is not always clear why particular imputation methods have been selected, with some notable exceptions (Lafontaine et al. 2024). In our analysis, we compared the model performance of common imputation methods by introducing synthetic missing data in sensitivity analyses. In total, 46.3% of the entire dataset was generated via imputation methods. Previous studies investigating nitrate contamination had between 18% and 87% of observation years imputed (Espejo-Herrera et al. 2016; Lafontaine et al. 2024; Schullehner et al. 2017; Ward et al. 2003; Weyer et al. 2001). Our results are consistent with previous assessments and contain a moderately high proportion of imputed values (Espejo-Herrera et al. 2016; Lafontaine et al. 2024; Schullehner et al. 2017; Ward et al. 2003; Weyer et al. 2001). As with previous studies, the historical under-sampling and issues with record keeping are persistent within NZ (Chambers, Hales, et al. 2022).

Our assessment spans 1964 to 2024; however, nitrate measurements prior to 1997 were sparse. From 1996 to 2004, there was a national sampling programme to assess potential drinking water risks that resulted in the observed spike in testing seen in 1997. From that point on, testing frequencies increased across NZ, with another spike around 2013. Consequently, with the inclusion of imputed data, the assessment from 2000 covers the majority (89.2%) of the available person-years of observation. Therefore, this dataset is more appropriate for examining health effects with shorter latency periods such as those associated with maternal or early childhood outcomes.

The distribution of nitrate concentrations across the population is very similar to a previous estimate (Richards et al. 2022). Our analysis suggested that 95.3% of the population on a council supply are exposed to nitrate concentration < 1 mg/L nitrate-nitrogen compared to 86.9% from a 2022 assessment (Richards et al. 2022). The main differences between these studies are that the current study includes slightly more recent data (2024 compared to 2021) and for more supplies (values missing for 18,000 people compared to 126,627). Further, there have been notable nitrate exceedances in three small supplies in the Canterbury (Pareora and Waimate) and Otago (Gore) regions, which are representative of wider nitrate contamination of freshwater in these areas (Prickett et al. 2023).

Internationally, population drinking water nitrate exposure distributions vary substantially. In France, national surveillance data indicate that whereas most of the population is supplied by distribution units with mean nitrate concentrations below approximately 5.6 mg/L nitrate-nitrogen, sizeable proportions experience higher levels, with around 15% exposed to mean concentrations between ~5.6 and 9.0 mg/L (Ministere De La Sante et de Lacces aux soins 2024). In Denmark, approximately 80.7% of the population is exposed to nitrate in drinking water at levels below 0.87 mg/L, 9.9% at levels between 0.87 and 2.09 mg/L, whereas a further 9.4% are at levels above 2.09 mg/L (Jacobsen et al. 2024). By comparison, our finding that more than 95% of the New Zealand population on a public supply have nitrate concentrations below 1 mg/L suggests a distribution strongly concentrated at the low end, which is supported by a low population-weighted median of 0.05 mg/L nitrate-nitrogen. This contrast likely reflects differences in water supply structure (surface v groundwater), and agricultural pressures, as well

as higher exposures associated with domestic self-supplies not being included in this current study.

4.1 | Strengths and Limitations

This study has several strengths. First, the dataset covers almost the entire population of New Zealand that receives water from a registered water supply. Second, the raw data used were from accredited laboratory results retrieved directly from either the water supplier (e.g., TAs) or water regulators (MoH and Taumata Arowai). Third, we implemented and validated a range of different methods to improve coverage including deriving values from source or treatment plant-level measurements, applying virtual zones and imputation methods, which were largely facilitated by documentation on the relationships between water supply components received directly from water regulators.

There were also limitations. Information on the relative contribution of each water supply component (source or plant) to a WDZ was not available, so we could not estimate a weighted average of the supply component values. Further, some supply components may have had a greater frequency of testing within a given year, which would lead to more accurate annual estimates for some sources/plants than others. The impact of this limitation is probably small given the marginal difference between the source- or plant-derived values and the WDZ values (see Supporting Information). Additionally, these minor differences extended to supplies with only a single source and treatment, which also suggests a proportion of the difference between values is likely due to differential timing of testing within the year between different supply components (e.g., source tested in March, WDZ tested in July).

There were difficulties with naming conventions for water supply components and unit measurements for nitrate (nitrate vs. nitrate-nitrogen). For the former, any supply component without a valid ID was manually identified using information from historical drinking water registries. TAs were also contacted for clarification on any remaining records with vague naming conventions (e.g., "bore 13"). Any unmatched values after this process were excluded.

There were several limitations with the imputation methods and analysis. First, our best performing model (XGBoost imputation) included a relatively modest number of predictor variables that did not include information on soil type, other forms of land use outside livestock numbers, or information on groundwater redox status. Further, the covariates included were at a relatively coarse spatial resolution (e.g., regional). Second, the imputation analysis assumes that the observed values in any given year are representative of the actual annual nitrate value. For example, most annual estimates in this dataset are derived from a single measurement. The supplementary results showed there was a preference to test in autumn, which had slightly lower nitrate values than in spring (0.16 mg/L nitrate-nitrogen) and winter (0.16 mg/L nitrate-nitrogen) across all water supply components. Of annual values relying on a single test that are most vulnerable to seasonal bias, 35% were conducted in autumn, 27% in summer, 23% in spring, and 15% in winter. These results suggest that if there is a seasonal bias in the data, it is likely towards slightly lower values (e.g., 0.16 mg/L on average).

There is an important limitation with generalizing from this study to estimate nitrate exposure for the whole population of NZ. The exposure estimates here could only include the registered water supplies that cover about 85% of the resident NZ population. The non-registered water supplies are largely small and often in more rural areas, so may have higher nitrate levels. A recent report from the Ministry for the Environment stated that "up to 10 per cent of unregistered supplies probably have median concentrations in excess of the nitrates MAV"(Ministry for the Environment 2022, 45).

4.2 | Future Improvements to the Surveillance of Drinking Water Quality

The limitations of nitrate exposure surveillance highlighted in the previous section should be addressed to ensure the system is better able to support its intended purpose of protecting public health. It is also important that the data are adequate to support longer-term assessment of trends in the level and distribution of nitrate exposure and research of the kind described in this report.

In part, some of these limitations have been addressed by the New Drinking Water Quality Assurance Rules, which came into effect in November 2022 and set out the frequency that drinking water from water supply components (source, treatment, and reticulation) must be monitored for key determinands, including nitrate (Taumata Arowai 2022). However, further improvements could include:

1. More frequent sampling for small supplies serving 26–100 people (currently once every 3 years) and median supplies serving 101–500 people (currently annually).
2. Monitoring at the WDZ level for supplies with two or more sources (as all current monitoring is done at the source) or information on the relative contribution from different sources to a particular water supply.
3. Strict implementation of a data standard to ensure data are collected and maintained in a standardized format. Data received from Taumata Arowai showed determinands were specified using free text (e.g., "nitrate from bore 2") and had inconsistent unit measurements (a combination of nitrate and nitrate-nitrogen), which were difficult to verify.
4. Appropriate sanctions for supplier noncompliance with the testing regime.
5. Systematic monitoring of a sample of nonregistered supplies to estimate exposures for the population served by these sources, which could be facilitated by Regional Councils responsible for source water protection and environmental regulation.

5 | Conclusions

This extensive assessment provides the foundation for epidemiological research into the health effects of nitrate contamination of drinking water in New Zealand. The effectiveness of

the system for drinking water nitrate surveillance could be enhanced in several ways that would improve its ability to meet its intended purpose.

Author Contributions

Tim Chambers: conceptualization, data curation, funding acquisition, methodology, project administration, methodology, validation, formal analysis, writing – original draft, writing – review and editing. **Frank Dean:** data curation, methodology, formal analysis, writing – review and editing. **Jacques Klavs:** data curation, methodology, formal analysis, writing – review and editing. **Nigel Stanger:** conceptualization, data curation, methodology, writing – original draft, writing – review and editing. **Alice Kim:** formal analysis, methodology, writing – review and editing. **Simon Hales:** conceptualization, data curation, formal analysis, funding acquisition, methodology, writing – original draft, writing – review and editing. **Jeroen Douwes:** conceptualization, funding acquisition, methodology, writing – review and editing. **Michael G. Baker:** conceptualization, funding acquisition, methodology, writing – review and editing. **Jeremiah Deng:** conceptualization, data curation, methodology, writing – original draft, writing – review and editing.

Acknowledgments

The authors would like to thank all the employees from the Territorial Authorities for providing the original data and information about their data systems. They would like to thank Chief Advisor Water Science Jim Graham from Taumata Arowai for his valuable advice throughout the project. We would also like to thank Yaser Dorgham for his initial work on the drinking water database. Open access publishing facilitated by University of Canterbury, as part of the Wiley - University of Canterbury agreement via the Council of Australasian University Librarians

Funding

The project was funded by the Health Research Council of New Zealand (HRC ref. 22/059). The funder did not have any input into the study design, collection, analysis and interpretation of data, writing of the report, and decision to submit the article for publication. The project was also supported with in-kind support from the Ministry for the Environment through staff time.

Ethics Statement

The project was approved by the University of Otago Human Ethics Committee (HD22/076).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The datasets generated during and/or analyzed during the current study are not publicly available but are available from the corresponding author on reasonable request.

References

Alam, S. M. K., P. Li, and M. Fida. 2024. "Groundwater Nitrate Pollution due to Excessive Use of N-Fertilizers in Rural Areas of Bangladesh: Pollution Status, Health Risk, Source Contribution, and Future Impacts." *Exposure and Health* 16, no. 1: 159–182. https://doi.org/10.1007/s12403-023-00545-0.

ANSES. 2022. "Reducing Dietary Exposure to Nitrites and Nitrates." ANSES. https://www.anses.fr/en/content/reducing-dietary-exposure-nitrites-and-nitrates.

Chambers, T., J. Douwes, A. t Mannetje, et al. 2022. "Nitrate in Drinking Water and Cancer Risk: The Biological Mechanism, Epidemiological Evidence and Future Research." *Australian and New Zealand Journal of Public Health* 46: 105–108. https://doi.org/10.1111/1753-6405.13222.

Chambers, T., S. Hales, N. Wilson, and M. Baker. 2022. "Improvements to Drinking Water Monitoring, Reporting and Record-Keeping Needed to Protect Health." *Policy Quarterly* 18: 23–27.

Espejo-Herrera, N., E. Gracia-Lavedan, E. Boldo, et al. 2016. "Colorectal Cancer Risk and Nitrate Exposure Through Drinking Water and Diet." *International Journal of Cancer* 139, no. 2: 334–346. https://doi.org/10.1002/ijc.30083.

Fida, M., P. Li, S. M. K. Alam, Y. Wang, A. Nsabimana, and P. S. Shrestha. 2024. "Review of Groundwater Nitrate Pollution From Municipal Landfill Leachates: Implications for Environmental and Human Health and Leachate Treatment Technologies." *Exposure and Health* 16, no. 5: 1225–1249. https://doi.org/10.1007/s12403-023-00624-2.

International Agency for Research on Cancer. 2010. "Ingested Nitrate and Nitrite, and Cyanobacterial Peptide Toxins (IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Issue)." https://publications.iarc.fr/_publications/media/download/2867/c9f9c85d6dd616d774bdbbe67bae77bddeb1b4de.pdf.

Jacobsen, B. H., B. Hansen, and J. Schullehner. 2024. "Health-Economic Valuation of Lowering Nitrate Standards in Drinking Water Related to Colorectal Cancer in Denmark." *Science of the Total Environment* 906. https://doi.org/10.1016/j.scitotenv.2023.167368.

Lafontaine, A., S. Lee, B. Jacquemin, et al. 2024. "Chronic Exposure to Drinking Water Nitrate and Trihalomethanes in the French CONSTANCES Cohort." *Environmental Research* 259: 119557. https://doi.org/10.1016/j.envres.2024.119557.

Land Air Water Aotearoa. 2023. "Factsheet: Nitrogen. Land Air Water Aotearoa." https://www.lawa.org.nz/learn/factsheets/nitrogen/.

Li, L., P. Li, S. He, D. Wang, Y. Tian, and L. Niu. 2025. "Groundwater Nitrate Pollution Source Apportionment Under Varying Land Use/Land Cover Patterns." *Exposure and Health* 17, no. 2: 353–369. https://doi.org/10.1007/s12403-024-00666-0.

Ministere De La Sante et de Lacces aux soins. 2024. "Bilan de La Qualite de L'Eau au Robinet du Consommateur Vis-a-Vis de Nitrates en France en 2023." Ministere De La Sante et de Lacces aux soins. https://sante.gouv.fr/IMG/pdf/bilan_qualite_nitrates_2023.pdf.

Ministry for the Environment. 2022. "Addressing Risks Associated With Nitrates in Drinking Water." Ministry for the Environment. https://environment.govt.nz/assets/publications/Freshwater/risks-associated-with-nitrates-in-drinking-water.pdf.

Prickett, M., T. Chambers, and S. Hales. 2023. "When the First Barrier Fails: Public Health and Policy Implications of Nitrate Contamination of a Municipal Drinking Water Source in Aotearoa New Zealand." *Australasian Journal of Water Resources* 28, no. 1: 64–73. https://doi.org/10.1080/13241583.2023.2272324.

Puente-Sierra, M., T. Chambers, L. Marek, J. M. Broadbent, B. O'Brien, and M. Hobbs. 2023. "The Development and Validation of A Nationwide Dataset of Water Distribution Zones in Aotearoa New Zealand: A Cross-Sectional Geospatial Study." *Data in Brief* 49: 109349. https://doi.org/10.1016/j.dib.2023.109349.

Richards, J., T. Chambers, S. Hales, et al. 2022. "Nitrate Contamination in Drinking Water and Colorectal Cancer: Exposure Assessment and Estimated Health Burden in New Zealand." *Environmental Research* 204: 112322. https://doi.org/10.1016/j.envres.2021.112322.

Royal, H., A. 't Mannetje, S. Hales, J. Douwes, M. Berry, and T. Chambers. 2024. "Nitrate in Drinking Water and Pregnancy Outcomes: A Narrative Review of Epidemiological Evidence and Proposed

Biological Mechanisms." *PLOS Water* 3, no. 1: e0000214. https://doi.org/10.1371/journal.pwat.0000214.

Schullehner, J., N. L. Jensen, M. Thygesen, B. Hansen, and T. Sigsgaard. 2017. "Drinking Water Nitrate Estimation at Household-Level in Danish Population-Based Long-Term Epidemiologic Studies." *Journal of Geochemical Exploration* 183: 178–186. https://doi.org/10.1016/j.gexplo.2017.03.006.

Taumata Arowai. 2022. "Drinking Water Quality Assurance Rules." Taumata Arowai. https://www.taumataarowai.govt.nz/assets/Uploads/Rules-and-standards/Drinking-Water-Quality-Assurance-Rules-2022-Released-25-July-2022.pdf.

United States Environmental Protection Agency. 2023. "Protocol for the Nitrate and Nitrite IRIS Assessment (Oral) (Preliminary Assessment Materials)." United States Environmental Protection Agency. https://www.regulations.gov/document/EPA-HQ-ORD-2017-0496-0010.

Ward, M. H., K. P. Cantor, D. Riley, S. Merkle, and C. F. Lynch. 2003. "Nitrate in Public Water Supplies and Risk of Bladder Cancer." *Epidemiology* 14, no. 2: 183–190. https://doi.org/10.1097/00001648-200303000-00012.

Ward, M. H., R. R. Jones, J. D. Brender, et al. 2018. "Drinking Water Nitrate and Human Health: An Updated Review." *International Journal of Environmental Research and Public Health* 15, no. 7: 1557. https://doi.org/10.3390/ijerph15071557.

Weyer, P. J., J. R. Cerhan, B. C. Kross, et al. 2001. "Municipal Drinking Water Nitrate Level and Cancer Risk in Older Women: The Iowa Women's Health Study." *Epidemiology* 12, no. 3: 327–338.

World Health Organization. 2016. "Nitrate and Nitrite in Drinking-Water (Background Document for Development of WHO Guidelines for Drinking-Water Quality, Issue)." World Health Organization. https://cdn.who.int/media/docs/default-source/wash-documents/wash-chemicals/nitrate-nitrite-background-jan17.pdf?sfvrsn=1c1e1502_4.

World Health Organization. 2017. "Guidelines for Drinking-Water Quality: Fourth Edition Incorporating the First Addendum." https://apps.who.int/iris/bitstream/handle/10665/254637/9789241549950-eng.pdf;jsessionid=653DBF3E25683F182CA7C095E5F61A04?sequence=1.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1:** Supporting Information. **Data S2:** Supporting Information. **Data S3:** Supporting Information.