Chronic exposure to drinking water nitrate and trihalomethanes in the French CONSTANCES cohort

Antoine Lafontaine, Sewon Lee, Bénédicte Jacquemin, Philippe Glorennec, Barbara Le Bot, Dominique Verrey, Marcel Goldberg, Marie Zins, Emeline Lequy, Cristina M. Villanueva

Environmental Research 259 (2024) 119557

Contents lists available at ScienceDirect

Environmental Research

journal homepage: www.elsevier.com/locate/envres

Chronic exposure to drinking water nitrate and trihalomethanes in the French CONSTANCES cohort

Antoine Lafontaine [a,*], Sewon Lee [b,c,d], Bénédicte Jacquemin [a], Philippe Glorennec [a], Barbara Le Bot [a], Dominique Verrey [a], Marcel Goldberg [e], Marie Zins [e], Emeline Lequy [e,**], Cristina M. Villanueva [b,c,d,f]

[a] Univ Rennes, Inserm, EHESP, Irset (Institut de Recherche en Santé, Environnement et Travail) - UMR_S, 1085, Rennes, France
[b] ISGlobal, Barcelona, Spain
[c] Universitat Pompeu Fabra (UPF), Barcelona, Spain
[d] CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain
[e] Unité "Cohortes en Population" UMS 011 Inserm/Université Paris Cité/Université Paris Saclay/UVSQ, Villejuif, France
[f] IMIM (Hospital del Mar Medical Research Institute), Barcelona, Spain

ARTICLE INFO

Keywords:
Disinfection by-products
Environmental monitoring
Exposure assessment
Water quality
Chemical safety
Public health

ABSTRACT

Trihalomethanes (THMs) and nitrate are widespread chemicals in drinking water. Chronic exposure has been associated with increased cancer risk despite inconclusive evidence, partly due to the challenges in long-term exposure assessment and potential exposure misclassification.

We estimated concentrations of nitrate and THMs in drinking water using a public regulatory monitoring database (SISE-Eaux) for CONSTANCES, a French population-based prospective cohort.

We obtained 26,322,366 measurements of drinking water parameters from 2000 to 2020. We excluded missing, implausible and duplicated measurements; we corrected or imputed missing geocodes of sampling locations; we calculated the annual median concentration of nitrate and THMs by surveillance area. To predict missing annual median concentrations, linear mixed models with random intercept using surveillance area as a clustering variable were developed for each region for nitrate and the four THM components (chloroform, chlorodibromomethane, bromodichloromethane and bromoform) separately. Concentrations in the nearest surveillance area from the household were merged per year among 75,462 participants with residential history geocoded for 2000–2020. Estimated concentrations resulting from this approach were compared with measured concentrations in 100 samples collected in Paris, Rennes and Saint-Brieuc in 2021.

Median annual concentrations of total THMs and nitrate at study participants' homes for 2000–2020 were, respectively, 15.7 μg/l (IQR: 15.2) and 15.2 mg/l (IQR: 20.8). Among these, 35% were based on measurements for nitrate (16% for THMs), 44% (46%) were predicted using on linear mixed models, and 21% (38%) were based on distribution unit median values. Conditional R^2 predictive models ranged from 0.71 to 0.91 (median: 0.85) for nitrate, and from 0.48 to 0.80 for THMs (median: 0.68).

These concentrations will allow future association analyses with risk of breast and colorectal cancer. Our cleaning process introduced here could be adapted to other large drinking water monitoring data.

1. Introduction

Disinfection by-products (DBPs) and nitrate are widespread chemicals in drinking water and constitute ubiquitous exposures in the population. DBPs are formed during drinking water treatment, by the reaction between disinfectants (e.g. chlorine) and organic precursors, bromide, or iodide naturally occurring in raw water. DBP formation further occurs along the distribution network through reaction of chlorine residual with organic matter, leading to increased levels in the tap water compared to the treatment plant (Rossman et al., 1994, 2001). Trihalomethanes (THMs) including chloroform, bromodichloromethane, dibromochloromethane, and bromoform are among the most

* Corresponding author.
** Corresponding author.
E-mail addresses: antoine.lafontaine@inserm.fr (A. Lafontaine), emeline.lequy-flahault@inserm.fr (E. Lequy).

https://doi.org/10.1016/j.envres.2024.119557
Received 26 April 2024; Received in revised form 13 June 2024; Accepted 2 July 2024
Available online 3 July 2024

abundant DBP classes and have been regulated in the EU since 1998 with a maximum contaminant level (MCL) of 100 μg/l (The European Parliament and the Council of the European Union, 2020). Chloroform and bromodichloromethane have been classified as possible human carcinogens by the WHO International Agency for Research on Cancer (IARC) (International Agency for Research on Cancer, 2012). THMs have been used as DBP surrogates in epidemiological studies and long-term exposure has been consistently associated with increased bladder cancer risk (Costet et al., 2011; Villanueva et al., 2004). However, evidence for other cancer sites such as colorectal (Helte et al., 2023; Rahman et al., 2010; Villanueva et al., 2017), breast (Font-Ribera et al., 2018; Koivusalo et al., 1997), or other (International Agency for Research on Cancer, 2012), is suggestive but remains mixed or insufficient.

Nitrate in drinking water mostly originates from the use of fertilisers in intensive agriculture and waste from intensive farming, and is regulated in the European Union (EU) since 1980 with a MCL of 50 mg/l (nitrate-ion) (The European Parliament and the Council of the European Union, 2020). Ingested nitrate has been classified by the WHO-IARC as a probable human carcinogen in conditions of endogenous nitrosation (International Agency for Research on Cancer, 2010). There is consistent evidence linking long-term exposure to nitrate in drinking water and increased colorectal cancer risk (Espejo-Herrera et al., 2016a; Schullehner et al., 2018; Ward et al., 2018) although a causal link has not been yet established (Elwood and van der Werf, 2022). Breast carcinogenicity of N-nitroso compounds (e.g. N-methyl-N-nitrosourea) was found in animal studies (Tsubura et al., 2011) and ingested nitrate in drinking water has been associated with breast cancer risk in postmenopausal women (Espejo-Herrera et al., 2016b), but evidence is limited (Inoue--Choi et al., 2012; Ward et al., 2018).

Evaluation of the links between drinking water contaminants and cancer risk is arduous (Villanueva et al., 2014). Study design has been mainly limited to case-control studies and the use of cohort design has been scarce. Exposure assessment needs to be retrospective in nature given that prospective studies would require long follow up periods, due to statistical power constraints, that may also be a source of participants loss. The lack of valid biomarkers of long-term exposure to nitrate or THMs forces the use of drinking water concentrations as personal exposure surrogates. Exposure assessment for an etiologically relevant period for cancer is challenging, as it requires to retrospectively estimate exposure during decades before cancer diagnosis. Limited availability of historical concentrations in drinking water and the lack of residential history of study participants are among the main challenges. The use of long-term centralised databases from routine monitoring of water quality is promising to estimate exposure to water pollutants for epidemiological purposes (Schullehner and Hansen, 2014). However, databases designed for surveillance are not directly fit for research purposes and may require cleaning procedures in order to adequately apply them for epidemiological research.

CONSTANCES is a large general population-based cohort in Metropolitan France including around 220,000 French adults aged 18–69 years at enrolment (2012–2020) after a random selection among beneficiaries of the French national health insurance and its affiliates, in 22 health screening centres in 20 French departments ("Constances", 2023; Zins et al., 2015). Information was collected at enrolment and annual follow-up by self-administered questionnaire, and by a health examination at enrolment and every four years, and include health, socio-demographic, occupational, environmental, and lifestyle factors. CONSTANCES cohort is also linked to administrative databases, including the French national health insurance database from which diagnosis of certain diseases, including cancers, are identified. In addition, participants' addresses have been collected and geocoded from enrolment, and for a subset of 80,600 participants, lifetime addresses prior enrolment were collected in 2020–2022. SISE-Eaux is the French national database for drinking water quality surveillance. Although several studies have previously used SISE-Eaux data, the cleaning methods have not been detailed (Corso et al., 2018; Tiouiouine et al., 2020). We describe here the data cleaning procedures of the SISE-Eaux database and the methods to estimate chronic concentrations of exposure to THMs and nitrate in the CONSTANCES cohort (Zins et al., 2015), along with results.

2. Material and methods

2.1. SISE-Eaux database

SISE-Eaux (*Système d'Information en Santé-Environnement sur les Eaux*) is a national drinking water monitoring database created in 1994, managed by the French Ministry of Health (Tiouiouine et al., 2020) and fed by the Regional Health Agencies (ARS-*Agence Régionale de Santé*). SISE-Eaux includes information about the concentration of regulated water quality parameters, sampling site and date, water source (ground, surface, mixed, or sea), and whether the sample is collected at the distribution network (e.g. tap) or at the treatment plant outlet. Sampling site information include administrative divisions (municipality and department), water distribution unit (UDI, *Unité de DIstribution d'eau potable*), surveillance area, geocode and location. Sampling site location is a manually written description of the site (e.g. "12 rue du Chêne" [12, Oak street], "Chez M. Dupont - Evier" [Mr Dupont's – kitchen sink], "Mairie" [City hall], etc.). In this study we focused on THMs, nitrate, and parameters known to be related to THMs and/or nitrate concentrations: conductivity, free and combined chlorine, total organic carbon, permanganate index, pH, and alkalinity (Chowdhury et al., 2008) for the period 2000–2020. In total, the database included 26,322,366 observations (Table 1), that were subject of a procedure to make data suitable for our research purposes (Fig. 1).

Measurement methods were not uniform over France and evolved over time. Nitrate concentrations were mostly determined by ionic chromatography (French norm 'NF EN ISO 10304–1'). Other methods

Table 1
Description of the available data for the following water quality parameters in this study's SISE-Eaux extraction before data cleaning, 2000–2020 period. Total observations (N), and percentage of missing values (NA's), measurements below the limit of detection (LOD), measurements in the treatment plant (TTP), and missing or abnormal geocodes by parameter. Source: French Health Ministry – Regional Health Agency (ARS)-SISE-Eaux.

Parameter	N	NA's (%)	<LOD (%)	TTP (%)	Missing and/or abnormal geocode (%)
Conductivity (μS/cm)	1,137,859	1.3	<0.1	19.0	84.3
Conductivity 25 °C (μS/cm)	4,327,287	0.1	0.2	21.7	83.0
Free chlorine (mg/l)	4,600,312	3.9	21.6	21.5	81.9
Total chlorine (mg/l)	4,127,098	4.0	17.1	21.5	81.3
Chlorate (μg/l)	2,450	26.6	31.1	53.3	83.1
Chlorite (mg/l)	64,195	14.3	49.2	31.9	86.8
Total organic carbon (mg/l)	945,545	0.3	16.7	86.8	71.7
Permanganate index[a] (mg O_2/l)	185,748	0.3	48.5	79.3	64.4
pH	6,070,205	0.4	<0.1	22.2	81.7
Alcalinity[b] (°f)	1,142,593	0.2	1.7	79.4	74.9
Chloroform (μg/l)	385,181	0.9	54.8	62.6	74.2
Bromodichloromethane (μg/l)	384,399	0.9	44.0	62.6	74.2
Dibromochloromethane (μg/l)	381,779	0.9	33.8	62.6	74.1
Bromoform (μg/l)	382,665	0.9	50.0	62.7	74.2
Nitrate (mg/l)	2,185,050	0.1	6.5	47.1	79.0
All	26,322,366	1.5	10.8	31.4	80.6

[a] Oxidation by potassium permanganate in acid medium under hot conditions.
[b] Measured as "titre alcalimétrique complet, TAC".

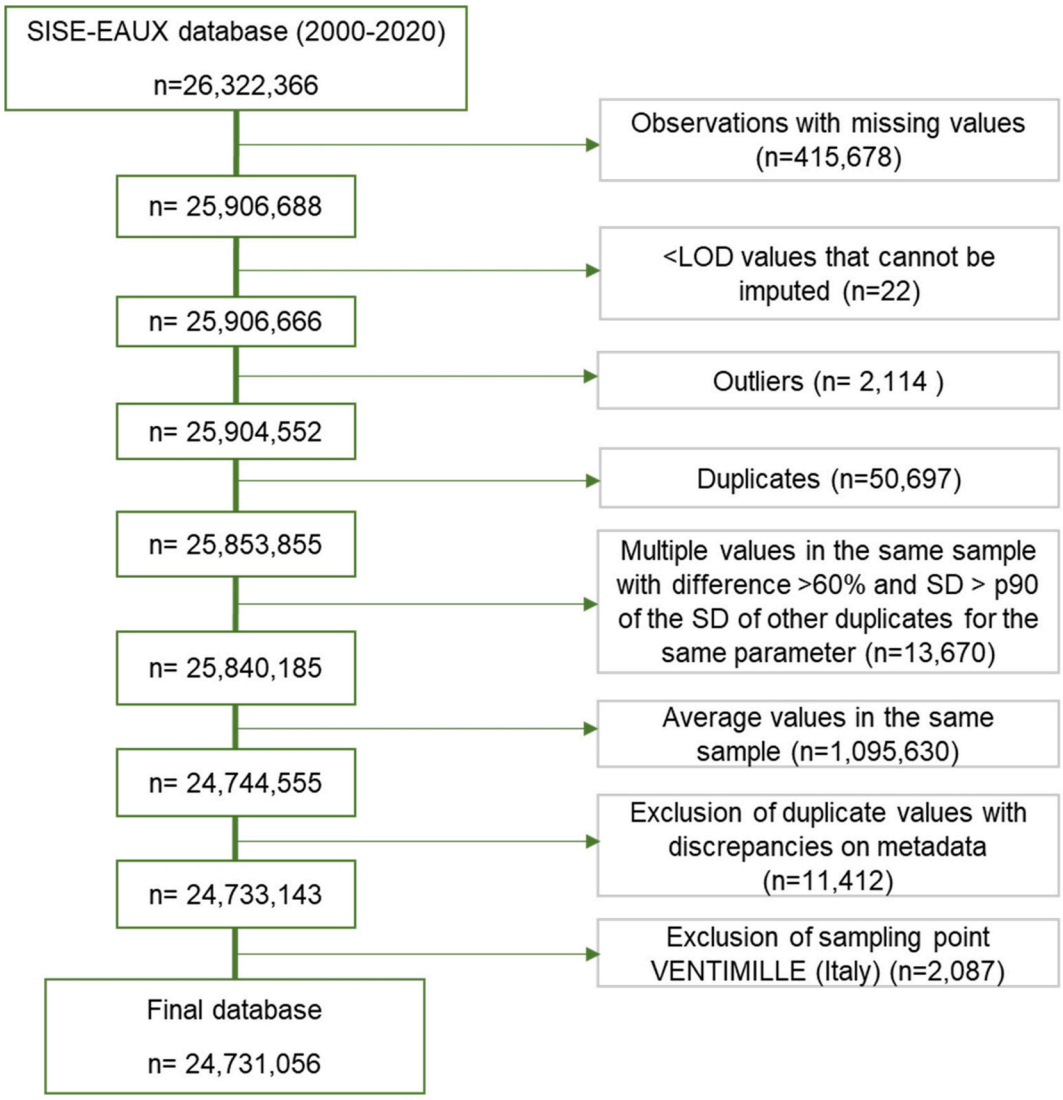


Fig. 1. Flow chart describing the sequential exclusion of observations of water quality parameters in SISE-Eaux for different reasons during the clean-up process. Source: French Health Ministry – Regional Health Agency (ARS)-SISE-Eaux.

were used, including continuous flow analysis and sequential method by colorimetry. THMs components were mostly determined by gas chromatography-mass spectrometry (GC-MS). GC-MS-MS was also marginally used in more recent years.

2.2. Cleaning and imputation for parameters with unknown concentrations

We excluded 415,678 observations as no concentration could be reasonably assigned (e.g. concentration coded as "?", "/", "N,M","NM") (Fig. 1). Concentrations coded as "PRESENCE", "<SEUIL", etc. were assumed to be below an unknown limit of detection (LOD) (N = 14,464). Values ≤ 0 (N = 248,156) or LODs 3 times higher than the 95th percentile of concentrations in the region (N = 2,752) were considered implausible and followed the same procedure to impute unknown LOD values based on known LODs. Given that LOD changed across time and space according to different equipment used (example shown in Table S1), we calculated the median LOD value by parameter, year, and department, and assigned this to unknown LOD values. If known LOD values were not available for the same year, the median LOD value from the closest year in the same department was assigned by K-Nearest-Neighbours (KNN) algorithm. In case of no known LODs for a given department, unknown LODs were imputed the median LOD from the same region and year. After this, values below LOD were imputed half the LOD for further calculations. Observations that were not possible to impute due to the absence of LOD in the same region were excluded (N = 22) (Fig. 1). A reduced number of measurements (<0.1%) were coded as '>' symbol followed by a number (ex: ">1.5"). These were imputed the specified value.

2.3. Outliers and implausible values

Outliers due to spelling errors (e.g. missing dots before the decimals of a number) were identified and manually corrected. Implausible values out of usual ranges according to the literature were identified for each parameter. pH values strictly lower than 4 or higher than 10 were excluded since such values were not plausible with the analytical methods used. We excluded values that exceeded twice the guidelines established by the World Health Organization (WHO) (World Health Organization, 2022) (Table S2). Nitrate levels in drinking water between 50 and 100 mg/l may be acceptable in France through a derogation procedure (ANSES, 2022). Hence, we considered 200 mg/l as a threshold to exclude outliers for this parameter instead of the WHO guideline. In total, 2,114 observations were excluded (Fig. 1).

2.4. Multiple measurements

Some parameters were measured twice in a few samples, corresponding to "field" and "laboratory" measurements. The "laboratory" measure was considered the most reliable value. However, information on the type of measure was not available and we followed other criteria based on the difference between the measures and their standard error. The standard error was used as a criterion to take into account that differences could be higher for values close to LODs. The maximum acceptable deviation for evaluating methods following the French norm 'NF T90-210' by doing an accuracy test comparing a concentration level measured to the level of estimated accuracy is 60% (Kinani et al., 2018). We chose this threshold for the difference between values. If two concentrations were identical, we kept one of the two (N = 50,697). When the difference between values was over 60% and the standard error of the two values exceeded the 90th percentile of the standard deviations of the other multiple observations for each parameter, the given multiple values were excluded (N = 13,670). Otherwise, we kept non-imputed concentrations (regarded as more reliable) over imputed ones. If the two concentrations were imputed, we kept the one imputed from a known LOD. Thus 11,254 observations were excluded. Samples with multiple measures with discrepancies on the measurement point (distribution network or treatment plant) and/or the location of the sampling site (e.g. different municipalities) were also excluded (N = 158 observations) (Fig. 1). All the other measures were averaged, leading to a total of 24,733,143 observations.

2.5. Harmonization of municipality codes

French municipalities may have changed over time: some being merged or divided. Municipality codes and changes are recorded by the National Institute of Statistics and Economic Studies (INSEE, *Institut National de la Statistique et des Études Économiques)*. Updates of INSEE codes have been tracked by searching the history of the codes on the INSEE's website. Each municipality was assigned one geocode representing its barycentre, extracted from the official database of postal codes, managed by 'La Poste' (2017) (https://www.data.gouv.fr/fr/datasets/base-officielle-des-codes-postaux/). Municipality surfaces were provided by the National Institute of Geographical and Forestry Information (IGN, *Institut National de l'information géographique et forestière*) in 2022 (https://www.data.gouv.fr/fr/datasets/correspondance-entre-les-codes-postaux-et-codes-insee-des-communes-francaises/).

2.6. Identification of abnormal geocodes

In order to check the plausibility of geocodes, we firstly projected the points on the map of France. Points that were projected outside of the map were considered as abnormal geocodes. For the remaining points, we computed the following metric:

$$m_p = \frac{dist_p{}^2}{surf\_area_p \times nbm_p}$$

where:

- $dist_p$ is the distance between a point p and the centroid of its municipality in kilometres
- $surf\_area_p$ is the surface area of its municipality in squared kilometres
- nbm_p is the number of municipalities with the same geocodes as the municipality of the point p.

Several municipalities had the same geocode since some centroids have not been updated after a dissolution of an older municipality or a merge of older municipalities. We took into account the number of municipalities with the same centroid geocode in order to penalize less points whose corresponding municipality surface has shrunk after a dissolution. All points with the previous metric above the 99th percentile of the distribution of the metric were considered as abnormal geocodes. Samples collected at the surveillance area *Ventimille*, in the Italian border were excluded (N = 2,087) given that samples corresponded to the Italian side.

2.7. Harmonization of geocode projection system

Abnormal geocodes appeared to be in projection systems other than Lambert93, the current official French projection system: Lambert1 (North), Lambert2 (*étendu* and *centre*), Lambert3, WGS84. Some geocodes had their X and Y coordinates inverted and/or had missing dots that were clearly identifiable. Other in WGS84 were misplaced in the character string detailing the location of the sampling. A total of 37.7% incorrect geocodes (51.9% of sampling points with tap water and 20.8% for treatment plants) could be corrected and converted into Lambert93 projection system (N = 31,039). The rest were considered as missing (N = 51,196).

2.8. Imputation of missing geocodes

Geocodes were available for some sampling locations at specific years and not others. The string character detailing the sampling site location was harmonised by text mining (removal of accents, putting everything into capital letters, removal of recurrent noise, etc.) and used to impute missing geocodes by a Last Observation Carried Forward (LOCF) approach grouped by surveillance area, measurement point (treatment plants or distribution unit), water source (ground, surface, sea, or mixed) and location. For instance, if we had the geocode for the location "Chez Monsieur Dupont – Evier Cuisine" and no geocode for "chez monsieur Dupont – Cuisine" in a same surveillance area with the same water source, both locations were transformed into "CHEZ MONSIEUR DUPONT" before the use of the LOCF approach.

Addresses were also found in the string character detailing the sampling location. In order to geocode these addresses we used the Address National Base (BAN, *Base Nationale Adresse*) through the National French platform 'adresse.data.gouv'. We included a filter on municipality codes during the geocoding to ensure that geocoded addresses would be placed in the correct municipality. For each input, we received geocodes, the corresponding addresses, the precision level of the geocoding (house number, locality, municipality or street) and a [0–1] precision score. We defined thresholds to exclude badly geocoded addresses for each level of precision: 0.30 for house number, 0.45 for street and 0.50 for locality. The thresholds were quite low since the geocoding tool was able to geocode correctly addresses containing remaining noise in the character strings (e.g. the substring "CHEZ MADAME DURAND –" in the location string "CHEZ MADAME DURAND – 21 RUE DES PEUPLIERS"). Kappa statistics were performed to assess agreement of the given addresses leading to 0.91 for house number, 0.84 for street, 0.77 for locality.

Finally, since city halls are widely used as sampling sites, we transformed string characters containing the word "MAIRIE" into just "MAIRIE" before applying again the LOCF approach grouped by surveillance point and water source to impute geocodes. This step was the last in order to avoid removing potential addresses for the previous geocoding method.

2.9. Data aggregation by surveillance area and year

In order to link concentrations with CONSTANCES participants' addresses, we used surveillance areas as the statistical unit. Surveillance areas are locations defined by the ARS to ensure the sanitary control of drinking water, and are considered to have a homogenous quality. Each surveillance area focuses either on distribution system/tap water

(NTTP) or treatment plants (TTP), and have different sampling frequency for each parameter. Measurements were conducted in different locations in a same surveillance area, month or year, and sampling frequency was not homogeneous (e.g. could be increased some months for different reasons such as pollution events). In order to have one concentration value by year, surveillance area, and parameter, we first aggregated the database per month, year and surveillance area and calculated the monthly median for each year. Missing water source was assigned the most frequent water source in the corresponding surveillance area. We then aggregated again the database per year and surveillance area to calculate the annual median concentration.

We geocoded surveillance areas using the centroid of existing sampling geocodes in the area. We computed distance between geocoded points in a same area: if the maximum distance between points was above 20 km and there were 5 or less geocoded points, we considered that the centroid of existing geocodes may not be correct and imputed it with the centroid of the municipality of the surveillance area. For surveillance areas with no geocodes, we also used the centroid of the municipality. Due to the lack of real geocodes, we had several surveillance areas allocated to the centroid of their municipality (e.g. the same point). Since the samples done in these areas could come from different water sources, we used a jittering approach instead of aggregating the values at the centroid: each component of the new geocodes (X and Y coordinates) were random values generated/extracted from a normal law with a mean equal to each component of the centroid of the municipality and a standard error of 100. We added rows for missing years for each surveillance area of the dataset. We used a last observation carried forward approach for the characteristics of the surveillance area (geocodes, source of water, region, department, UDI). Concentrations of these generated rows were considered as missing. The aggregated data is summarised in Table S3.

2.10. Estimated concentration of THMs and nitrate in surveillances areas

In order to estimate missing annual THM and nitrate median levels by surveillance area (Table S3), we used linear mixed models for each chemical: nitrate, chloroform, bromodichloromethane, dibromochloromethane, bromoform, and total THMs (sum of chloroform, bromodichloromethane, dibromochloromethane, and bromoform) adjusted for water source, department, and year, with random intercept with surveillance area as a clustering variable. We tested different models by region using different transformations in parameters depending on their distributions: square root and log-transformation. In order to select the best model, we computed the conditional R^2, e.g. proportion of total variance explained through both fixed and random effects. We then selected the model with the highest R^2. In some cases, when the assumptions of the model with a highest R^2 did not seem correct and the ones of the model with the second highest R^2 seemed better, we selected the second model. An interaction term between year and water source was added to include source-time trends in the models. Year was used as a factor variable since temporal trends were rarely linear. Separate models were run for measurements at TTP and at NTTP.

Two prediction sets were established: a marginal prediction which only uses the fixed part of the model, and a conditional prediction which uses both (fixed and random). Conditional predictions could only be conducted for surveillance areas used in the model. We predicted missing values of THMs and nitrate for each model and retained the conditional predictions. Negative predicted values were imputed half the median LOD of the corresponding parameter, region and year. When concentrations were missing for a surveillance area belonging to an UDI with data from other surveillance areas, we calculated the annual UDI median values to impute missing parameters given that water quality is considered homogeneous within UDI. Values that could not be predicted or imputed remained as missing. Different aspects of these choices will be discussed later on.

The inclusion of other parameters as covariates reduced the number of observations in the models due to missing values, and we only compared conditional R^2 of models with or without the following parameters: pH, conductivity and free chlorine. As a previous step, the two conductivity parameters available (Table 1) were merged into one by applying a conversion factor and missing free chlorine was imputed with total chorine in order to maximise the number of observations.

2.11. Estimated individual concentration of exposure of CONSTANCES participants

The study population included 75,462 CONSTANCES participants with geocoded residential history available for the 2000–2020 period. Exposure assigned was exclusively based on measurements in samples from surveillance areas defined by NTTP, that are closer to the consumers compared to TTP. To each non-missing geocode (n = 1,646,819) of CONSTANCES participants, we linked the water parameter's concentration of the nearest surveillance area from the residence. We did so even with missing exposure data to minimise misclassification by ensuring that a participant would not be linked to any point other than the nearest. For missing geocodes in Metropolitan France (n = 4,533), and geocodes from foreign countries (n = 20,997) for CONSTANCES participants we could not assign concentrations, and thus considered them as missing. If a participant lived at two or more addresses in the same year, the time spent at each address was used to compute a weighted mean.

2.12. Comparison between estimated and measured concentrations

A sampling campaign was conducted in Paris, Rennes and Saint-Brieuc in September 2021 in public places (e.g. schools, hospitals, restaurants, hostels). Nitrate, chloroform, bromodichloromethane, dibromochloromethane, bromoform, and total THMs were measured in drinking water samples in 50 sampling points in Paris, 25 in Rennes, and 25 in Saint-Brieuc (total 100 samples). We considered these sampling points as residential addresses of fictitious individuals and used our approach to estimate concentrations at these points using SISE-Eaux data for 2021. THMs concentrations in 2021 were not available for Paris. We examined the Pearson correlation coefficients between the annual residential concentrations obtained with the SISE-Eaux database in 2021 to the residential concentrations obtained with the sampling campaign.

All analyses were performed using R version 4.0.4 (R Foundation for Statistical Computing).

3. Results

3.1. Description of the SISE-Eaux database

This study's SISE-Eaux extraction contained 26,322,366 observations before cleaning, corresponding to different parameters (Table 1). On average, 31% of observations corresponded to measurements in TTP and the rest corresponded to samples in NTTP. Chloroform and bromoform showed the highest proportion of values below the LOD, 55% and 50%, respectively. In total, 80.6% of sampling locations were poorly or wrongly geocoded. Geocoding quality differed by region, from 7.9% of the sampling locations correctly geocoded in *Auvergne Rhône Alpes* region to 57.1% in *Pays de la Loire* region. After cleaning and imputation of missing and incorrect coordinates, 9 regions out of 13 had more than 40% surveillance areas correctly geocoded (87.3% and 92.1% for the *Provence Alpes Côte d'Azur* and *Centre Val de Loire* regions, respectively). On the contrary, the *Corse* region had 67.4% of its surveillance areas allocated to the barycentre of their municipalities.

Sampling frequency differed by surveillance area and parameter. For instance, 53.2% of the surveillance areas had five or less annual median concentrations of nitrate, and 83.9% for bromoform. Monitoring frequency increased over the years, from 14.8% of surveillance areas

analysing nitrate in 2000 to around 24% in 2020, and from 0.2% of surveillance areas analysing chloroform in 2000 to 11.2% in 2020. This increase is particularly important for THMs between 2003 and 2004 (0.8% of surveillance areas in 2003 and 4.4% in 2004).

After aggregation of the database, water source was 77.6% ground, 13.6% surface, and 8.7% mixed. Sea and missing accounted for less than 0.1%. These proportions differed by region. For instance, Brittany showed the highest surface water (45.6%) and the lowest percentage of ground water (19.1%) sources. Approximately 40% and 23% of surveillance areas had only one chloroform and nitrate value, respectively, after aggregation by year. Four regions (out of 13) (Fig. S3) had <70% missing annual nitrate concentrations. Brittany had the highest percentage of annual nitrate data (69.3%), followed by Normandy (48.8%), *Centre Val de Loire* (47.2%) and *Hauts de France* (30.9%). For chloroform, only two regions have more than 10% of annual concentrations.

3.2. Trihalomethane and nitrate concentrations

Temporal trends in concentrations differed across regions and by water source. In Fig. 2 we show the case of nitrate in Brittany and *Ile-de-France* as examples. In Brittany, there is a decreasing pattern of nitrate levels for all water sources (Fig. 2A). Those in sea water were extremely low compared to the others. In *Ile-de-France*, annual concentrations in ground water showed a slightly decreasing pattern over the years while levels in mixed and surface water decreased between 2003 and 2004 to rise again until 2020 (Fig. 2B). Nitrate levels remained stable over time

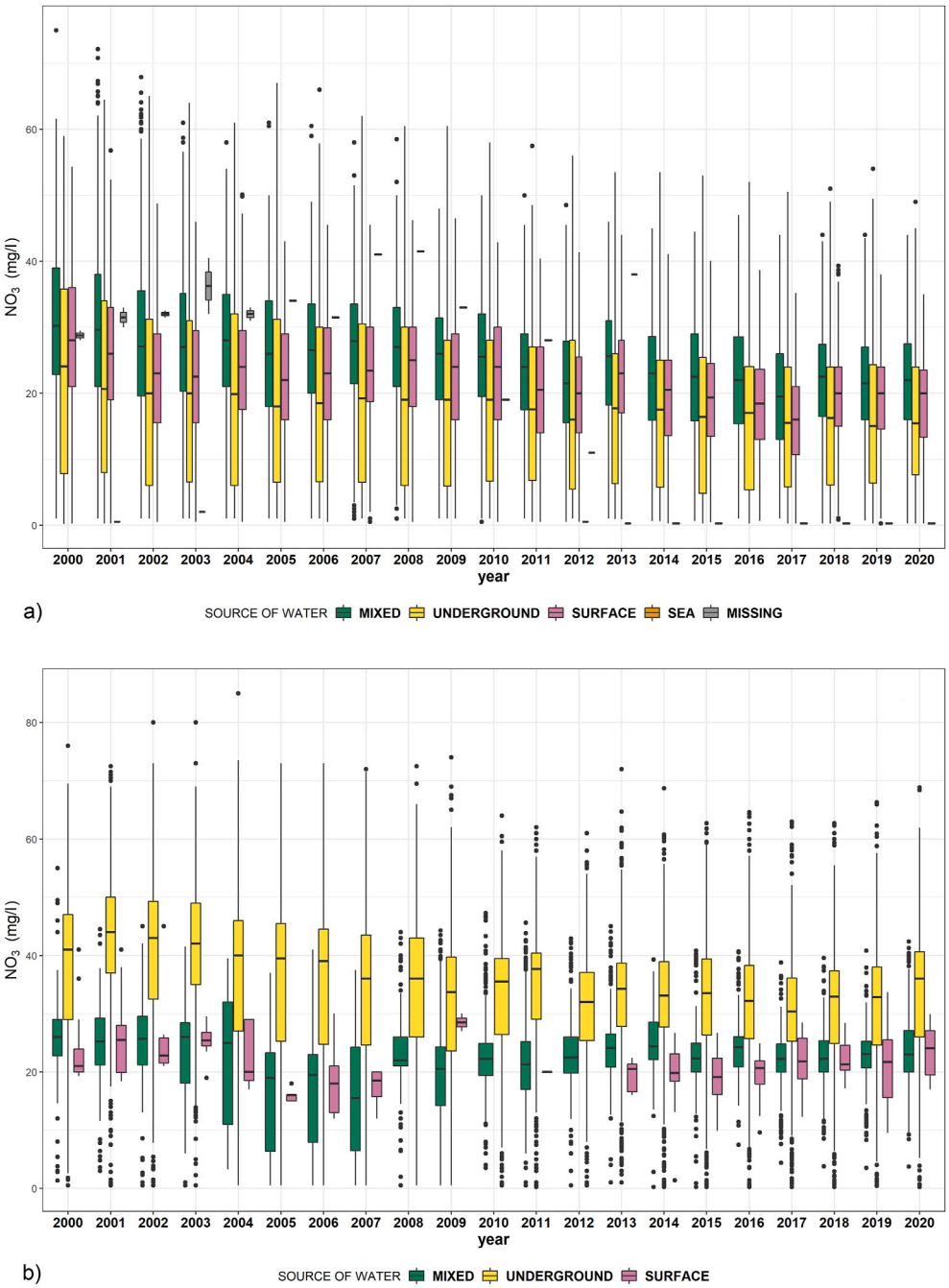


Fig. 2. Example of annual nitrate concentrations distribution in Brittany (a) and Ile de France (b) regions by water source. The line within the box indicates the median concentration and the boundaries of the box indicate the 25th to the 75th percentiles. Source: French Health Ministry – Regional Health Agency (ARS)-SISE-Eaux.

in other regions. Annual levels of individual THMs were below the WHO guidelines in all regions, except in a few surveillance areas for some years. Bromodichloromethane concentrations were above WHO guidelines in surface water in 2001. In general, chloroform, bromodichloromethane, dibromochloromethane, and bromoform concentrations were higher in surface than in ground water. THM concentrations were usually higher before 2004.

3.3. Multivariate models used to impute missing trihalomethanes and nitrate annual average concentrations

Conditional R^2 were higher than marginal R^2 in all models, meaning that most of the variation in the concentrations is explained by the random part of the model e.g. surveillance areas, the spatial effect. Conditional R^2 were particularly high for nitrate, from 0.71 to 0.91 (median: 0.85) for NTTP. For THMs, conditional R^2 were higher in *Auvergne Rhône Alpes* and lower in Brittany (0.80 and 0.53 respectively). Adding covariates (pH, conductivity, and free chlorine) did not change the models substantially in terms of marginal R^2: 0.80 with and without covariates for nitrate and 0.49 with or without covariates for bromoform in Brittany for NTTP (0.91 and 0.91 for nitrate and 0.79 and 0.78 for bromoform in the *Auvergne Rhône Alpes* region). As no prediction was done for surveillance areas without any annual concentrations, imputing the median of the UDI allowed us to impute more than 30% of the chloroform values of each region, except for Brittany where most of the values have been predicted. Even after predicting and imputing values, 46.7% of the THMs values of the *Grand Est* region remained missing (see Fig. 3 and Table S4).

3.4. Personal exposure information among CONSTANCES participants

The median (interquartile range) distance from the participants' residences to the nearest surveillance area was 451 (651) meters. The residential geocode was based on the exact address for 63% study participants, street-match level for 26%, the neighbourhood centre (as defined by IRIS - *Ilots Regroupés pour l'Information Statistique*) in 6%, and postal code centre in 5%. In total, 85% nitrate and total THM annual values were based on real sampling geocodes, and the rest were imputed based on the centroid of several geocodes available for the same sampling area. A total of 44% of nitrate (46% of total THM) annual values were based on model predictions, 34% (16%) were based on actual concentrations, and 22% (38%) were based on UDI median values. Median concentration of chloroform, bromodichloromethane, dibromochloromethane, bromoform, total THMs, and nitrate in the residence of study participants for the period 2000–2020 was, respectively, 1.8, 2.8, 4.8, 3.3, 15.6 µg/l, and 15.3 mg/l. Around 99% of THMs and nitrate concentrations at this study participants' home were below the WHO guidelines (sum of ratios "THM component values: guideline values" ≤ 1 and 50 mg NO_3/l) for the period 2000–2020, with higher levels in specific areas. Fig. 4 depicts the annual nitrate and total THM concentration among study participants for the period 2000–2020. None of the participants were allocated to a surveillance area dealing with sea water.

3.5. Comparison between estimated and measured concentrations

Nitrate concentrations in our sampling and those based on the SISE-Eaux database were highly correlated in Paris ($r = 0.85$, $p < 0.001$). In Saint-Brieuc, correlations were moderate to low (from $r = -0.11$ for nitrate to $r = 0.45$ for bromodichloromethane, all p-values were above 0.05) and distributions of the different parameters showed no contrast between UDIs (Fig. 4, Fig. S1). In Rennes, correlations ranged from $r = -0.26$ for nitrate to $r = 0.39$ for total THMs, with all p-values above 0.05. For all cities, contrasts between UDIs in the two databases were similar (see Fig. 5). Concentrations of THMs were higher in our sampling than those in the SISE-Eaux database as opposed to nitrate.

4. Discussion

We described the process to repurpose a database originally created for routine monitoring of drinking water quality in order to assess exposure in epidemiological research. We focused on THMs and nitrate in Metropolitan France, and aimed to cover an exposure period

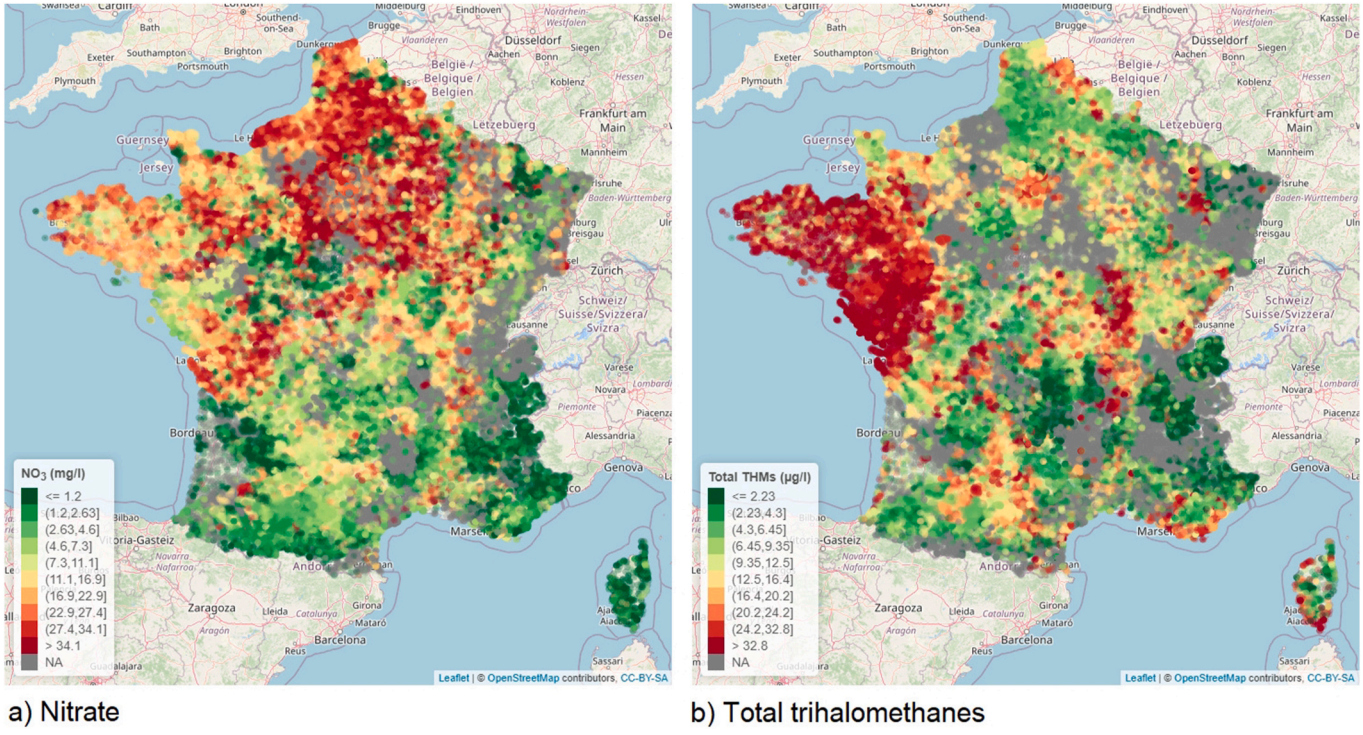


Fig. 3. Maps of the median concentrations of total THMs and nitrate by surveillance area in 2020 in mainland France (SISE-Eaux) after prediction and imputation. Source: French Health Ministry – Regional Health Agency (ARS)-SISE-Eaux.

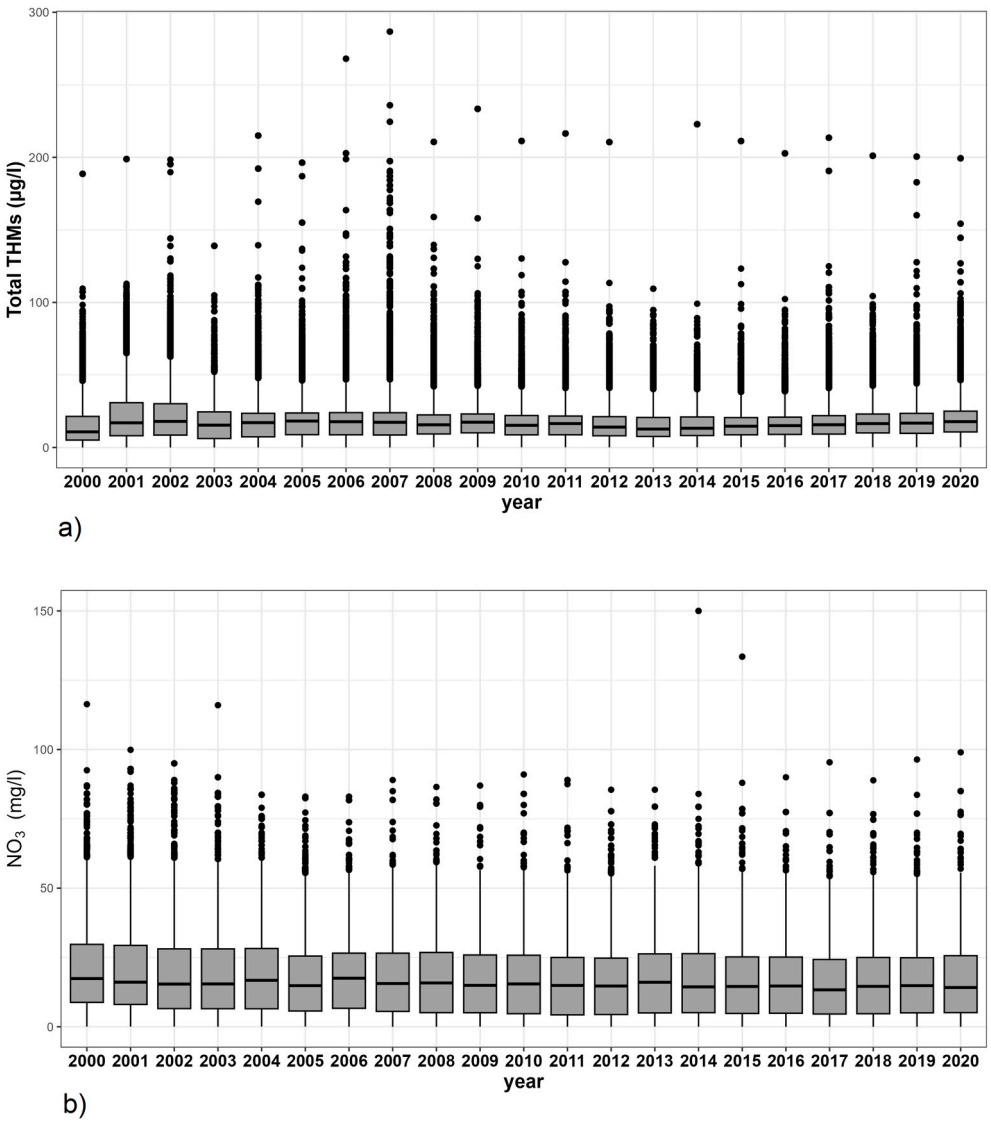
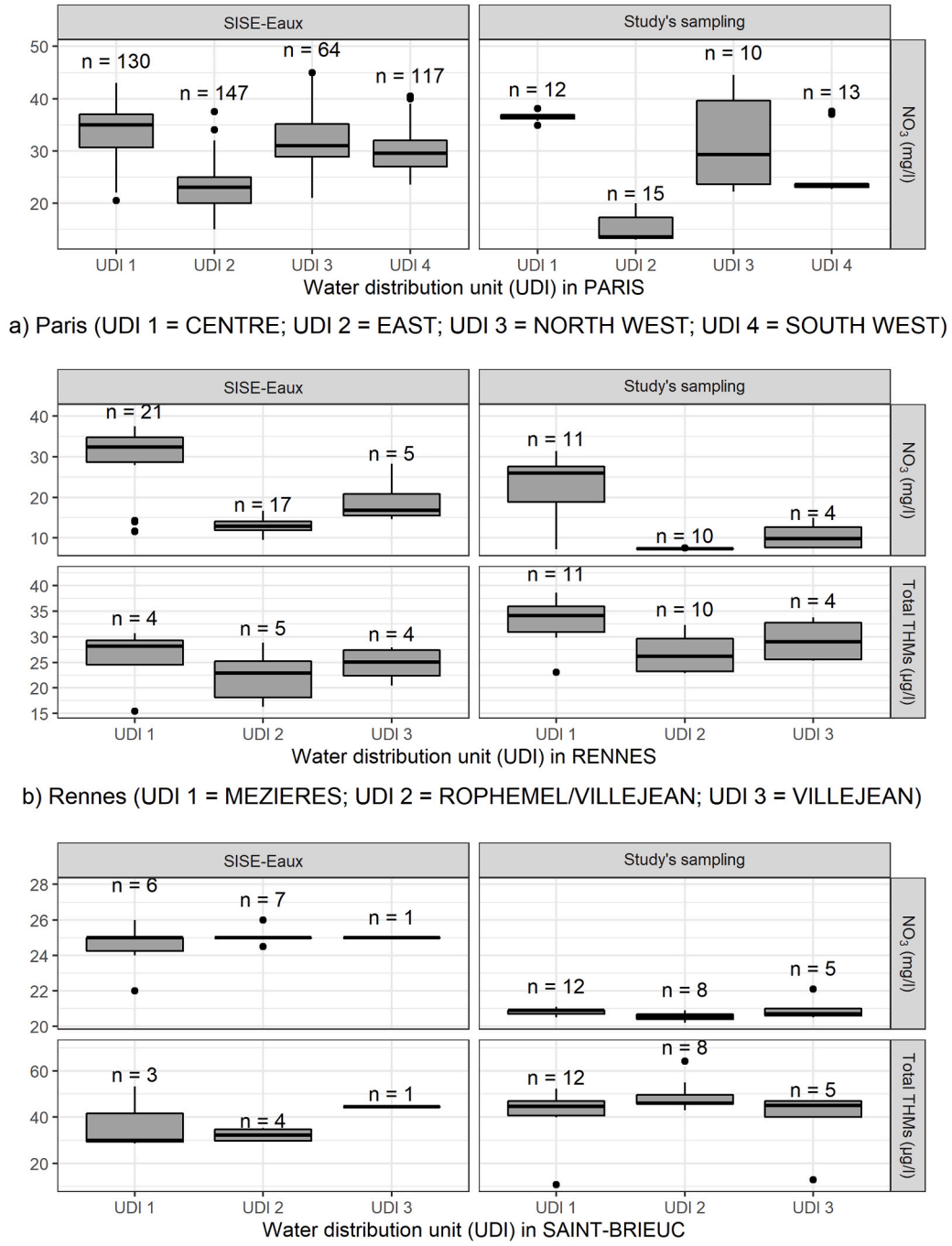


Fig. 4. Annual residential concentrations of total THMs (a) and nitrate (b) among CONSTANCES participants. The line within the box indicates the median concentration and the boundaries of the box indicate the 25th to the 75th percentiles. Source: French Health Ministry – Regional Health Agency (ARS)-SISE-Eaux.

etiologically relevant for cancer in the framework of the prospective CONSTANCES cohort study in France. This study addresses key challenges such as data cleaning of parameters of interest in drinking water and geocoding sampling points in a large dataset. Estimated concentrations resulting from this approach were compared with measured concentrations in collected samples showing various results depending on the number of points and the geocoding quality. We linked the water parameter concentration of the surveillance area nearest to the CONSTANCES participants residences to created individual exposure estimates.

Data availability on water quality varied over time and across geography. Sampling frequency depended on population density, and ranged from once every 10 years in areas with less than 50 inhabitants served to once a month for more than 300,000 inhabitants served. Samples were more frequently collected at the outlet of TTP than at the tap level. On the other hand, there are more surveillance areas corresponding to NTTP compared to TTP. Specifically, THMs were only routinely analysed since 2004, after the 2003 French regulation following the European directive 98/83/CE of November 1998 (Mouly et al., 2009). Sampling frequency has been more stable since 2005 (Corso et al., 2017). These different sampling frequencies have an impact on our predictions. First, estimated THMs concentrations before 2004 are less accurate and reliable since we do not know the reason of the sampling, that might be related to hotspots. Interpretation of temporal trends must remain cautious. Post-chlorination was always applied to surface water but was only promoted in 2003 by health authorities for groundwater and implemented progressively (Corso et al., 2018). This could seem contradictory with the fact that THMs values were higher in tap water coming from ground water for some regions before 2003 and a few years later (e.g. bromoform in tap water in Normandy or *Occitanie*). Nitrate concentrations were found to be above the WHO guideline (50 mg/l) in most of the regions, but this occurred in specific surveillance areas. These higher levels were local (agricultural zones in the south-west of Paris for instance). Despite intense agricultural activity, Brittany managed to lower the level of nitrate in water below the WHO guideline (50 mg/l) in 2020. Overall, the mean estimated nitrate concentration was lower than the mean concentration found in groundwater in Europe (16.4 mg/l in our population for 2000–2020 vs. 21.0 mg/l in Europe for the same period) (European Environment Agency, 2023). Estimated THMs concentrations seemed to be higher than those in Europe (mean of 19.1 µg/l in our population for 2000–2020 vs. population weighted mean of 11.7 µg/l in Europe for the 2005–2018 period) (Evlampidou et al., 2020). However, accuracy and coverage were not homogenous among countries in this study and mean



a) Paris (UDI 1 = CENTRE; UDI 2 = EAST; UDI 3 = NORTH WEST; UDI 4 = SOUTH WEST)

b) Rennes (UDI 1 = MEZIERES; UDI 2 = ROPHEMEL/VILLEJEAN; UDI 3 = VILLEJEAN)

c) Saint-Brieuc (UDI 1 = RESERVOIR BERRIEN; UDI 2 = MANŒUVRE; UDI 3 = VILLE HELIO)

Fig. 5. Comparison of nitrate (NO_3) and total THMs concentrations between this study's sampling campaign (measured) and SISE-Eaux (estimated) by distribution unit (UDI) in Paris, Rennes, and Saint-Brieuc. The line within the box indicates the median concentration and the boundaries of the box indicate the 25th to the 75th percentiles. Source SISE-Eaux data: French Health Ministry – Regional Health Agency (ARS)-SISE-Eaux.

concentrations were based on treatment plant concentrations for some, including for France.

Concentrations were considered homogeneous in a same surveillance area although measurements were not always conducted at the same spot. THM levels are known to vary seasonally (Charisiadis et al., 2015; Xu et al., 2022). Given that we aimed to generate an exposure metric etiologically relevant for chronic diseases, we prioritised to evaluate annual averages by-passing seasonal trends. Annual aggregated value in a surveillance area represented the median of a monthly-varying number of measurements. In that case, the representativity of the annual median value depended on the number and distribution of samples over the months (Mouly et al., 2009). Total organic matter and pH are also water quality parameters often used to predict THMs concentrations (Brown et al., 2011; Hong et al., 2020; Uyak et al., 2005; Xu et al., 2022). However, total organic carbon was sampled essentially in treatment plants and not at the tap level in France and we were not able to make use of this variable.

During the data cleaning process of SISE-Eaux database, we dealt with spelling mistakes, typing errors, or missing decimal points. Several outliers appeared to be measurements recorded in the wrong unit. These

errors tended to happen less frequently in recent years but correcting these values manually is essential when historical data is used in retrospective studies. Procedures for handling outliers are not usually provided, especially when individual exposures considering tap water consumption are reported (Menard et al., 2008). Moreover, previous studies did not use the same measurement window or spatial statistical individual for the estimation of concentrations of parameters as we did in this study (Costet et al., 2012). Therefore, it was challenging to have an insight on how researchers manage errors in raw monitoring data. Most studies using monitoring data rarely provided details on their data cleaning process as well as the study using the SISE-Eaux database (Beaudeau et al., 2010).

Correcting geocodes of sampling points presented another challenge. Many geocodes were missing or belonged to different map projection systems (French Lambert 93 or WGS 84, for instance), which could induce misplaced sampling sites leading to exposure misclassification. Before feeding nation-wide databases, effort should be made to harmonize local databases, especially regarding the map projection systems, and users of such databases must first carefully check them. Ideally, geocodes should be accurately recorded for each sample or at least for each surveillance area. We also found different LODs for a same parameter by department and year. Even if different equipment was used over time and its accuracy increased in recent years, we also found different LODs for a same parameter, department and year. Laboratories may have equipment with different performance.

Three potential statistical units could have been used to assign annual median concentrations to CONSTANCES participants. First, assessment could have been done based on the UDI. Since levels of parameters are homogeneous within UDI, using geocodes would not have been necessary in this case. Unfortunately, UDIs keep changing over time due to merging or disaggregation and there is no history of UDI geographic vector data covering all France and all years, so this approach was not possible. Secondly, sampling site could have been used. However, with nothing to identify them easily except their geocode, it was hard to determine whether several very close sampling points represented multiple sampling sites or a single sampling site which was poorly geocoded over time. Modelling at the sampling site was complicated, since there were too few repeated measures by sampling site. For these reasons, our choice was the surveillance area, represented by a point, since no polygon was available to delineate such areas. As a result, annual concentration assessment has to be done by taking the water quality value of sampling area (as a point) closest to each participant's residence. With a total of 134,057 surveillance areas in mainland France, this approach allowed us to get a good contrast of concentrations across the country.

Exposure measurement error is expected to be higher for trihalomethanes compared to nitrate given that there were more measurements for nitrate than for THMs, and also predictive models performed better (overall median R^2 was 0.85 for nitrate vs. 0.68 for total THMs). In total, 35% annual nitrate estimates assigned to study subjects corresponded to actual concentrations (vs. 16% of total THMs estimates). In addition, exposure misclassification can be higher in less-densely populated areas with fewer surveillance areas and poorly geocoded areas since we have solely one point representing an entire area. Our approach may be limited when dealing with peculiar UDI shapes, where the closest surveillance area does not necessarily belong to the UDI providing water to a participant's residence. Areas with a high number of well-geocoded points let us capture the shapes of these UDIs and to reduce the risk to assign a participant to a wrong UDI. Indeed, correlations between real concentrations and estimated concentrations of nitrate were more correlated in Paris than in Rennes or Saint-Brieuc, due to its higher number of correctly geocoded sampling areas. It is remarkable that we collected samples in September 2021, one month after a period of drought in Brittany, while SISE-Eaux data included several months. The limited comparability of timespan can additionally explain the moderate to low correlations between both measurements in Saint-Brieuc and Rennes. Indeed, punctual pics in temperature are smoothed by taking annual concentrations with SISE-Eaux.

There are very limited previous examples of nation-wide studies evaluating exhaustively concentrations of drinking water contaminants. To our knowledge, our study is only comparable to the previous Danish study that evaluated nitrate exposure from drinking water over 35 years using routine monitoring data (Schullehner and Hansen, 2014). Still, there are remarkable differences in terms of population (5 million in Denmark vs. approximately 65 million inhabitants in Metropolitan France), number of municipalities (98 in Denmark, 34,826 in Metropolitan France), water supply areas (2852 in Denmark, 134,057 surveillance areas in Metropolitan France) and water source (virtually 100% ground in Denmark, multiple sources in France). In addition, we included THMs while the Danish study focused only on nitrate. Both studies set a precedent that we hope will be expanded to other countries to exploit nation-wide centralised water quality data for epidemiological research despite the inherent challenges.

5. Conclusion

We described in detail the procedure of SISE-Eaux database cleaning, as a first step to estimate long-term exposure concentrations to THMs and nitrate in the CONSTANCES cohort. Future epidemiological analysis based on our nitrate and THM exposure estimates should consider accuracy in balance with statistical power. The cleaning process introduced here could be adapted to other large drinking water monitoring data in future studies.

Ethical approval

The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. All procedures have been approved by the Institutional review board (IRB) of the French Institute of Health (Inserm) (Opinion n°01–011, then n°21–842), and authorized by the by the French Data Protection Authority ("*Commission Nationale de l'Informatique et des Libertés*", CNIL) (Authorization #910486). The biobank obtained a favorable opinion from the Committee for the protection of individuals – CPP Sud Est I (#2018–32) and an authorization from the CNIL (#DR-2-2018-137). All volunteers sign a written consent form for their participation in CONSTANCES, and, where applicable, for their participation in the biobank. All information and regulatory authorizations relating to the publication are available on the French version page 'Rights and data protection'.

Funding

This project has received funding from the ANSES programme under Agreement num. 2019/1/049 - *Association entre exposition aux polluants dans l'eau de boisson et cancers du sein et du colon* (CancerWatch). The CONSTANCES cohort study was supported and funded by the French national health insurance fund ("Caisse nationale d'assurance maladie", Cnam). CONSTANCES is a National infrastructure for biology and health ("Infrastructure nationale en biologie et santé") and benefits from a grant from the French national agency for research (ANR-11-INBS-0002). CONSTANCES is also partly funded to a small extent by industrial companies, notably in the healthcare sector, within the framework of Public-Private Partnerships (PPP). None of these funding sources had any role in the design of the study, collection and analysis of data or decision to publish.

CRediT authorship contribution statement

Antoine Lafontaine: Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization.

Sewon Lee: Writing – original draft, Formal analysis. **Bénédicte Jacquemin:** Writing – review & editing, Project administration, Methodology, Funding acquisition, Conceptualization. **Philippe Glorennec:** Writing – review & editing, Resources. **Barbara Le Bot:** Writing – review & editing, Resources. **Dominique Verrey:** Writing – review & editing, Resources. **Marcel Goldberg:** Writing – review & editing, Resources, Investigation, Funding acquisition. **Marie Zins:** Writing – review & editing, Resources, Investigation, Funding acquisition. **Emeline Lequy:** Writing – review & editing, Methodology, Conceptualization. **Cristina M. Villanueva:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgements

We acknowledge the valuable contribution of the *Direction Générale de Santé* (DGS), Regional Health Agencies of *Île-de-France* and Britanny in providing SISE-Eaux data, and Eunate Abilleira Cillero and Enrique Ulibarrena (Public Health Laboratory in Gipuzkoa, San Sebastián, Spain) for the nitrate measurements in the collected drinking water samples. The authors thank the team of the "Population-based cohorts unit" (*Cohortes en population*) that designed and manages the CONSTANCES cohort study. They also thank the French national health insurance fund ("*Caisse nationale d'assurance maladie*", Cnam) and its Health screening centres ("*Centres d'examens de santé*"), which are collecting a large part of the data, as well as the French national old-age insurance fund ("*Caisse nationale d'assurance vieillesse*", Cnav) for its contribution to the constitution of the cohort, and ClinSearch, Asqualab and Eurocell, which are conducting the data quality control.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envres.2024.119557.

References

ANSES, 2022. ANSES Opinion and Report on the Assessment of Risks Associated with the Consumption of Nitrates and Nitrites. Maisons-Alfort: ANSES. Request No 2020-SA-0106).

Beaudeau, P., Valdes, D., Mouly, D., Stempfelet, M., Seux, R., 2010. Natural and technical factors in faecal contamination incidents of drinking water in small distribution networks, France, 2003–2004: a geographical study. J. Water Health 8, 20–34. https://doi.org/10.2166/wh.2009.043.

Brown, D., Bridgeman, J., West, J.R., 2011. Predicting chlorine decay and THM formation in water supply systems. Rev. Environ. Sci. Biotechnol. 10, 79–99. https://doi.org/10.1007/s11157-011-9229-8.

Charisiadis, P., Andra, S.S., Makris, K.C., Christophi, C.A., Skarlatos, D., Vamvakousis, V., Kargaki, S., Stephanou, E.G., 2015. Spatial and seasonal variability of tap water disinfection by-products within distribution pipe networks. Sci. Total Environ. 506, 26–35. https://doi.org/10.1016/j.scitotenv.2014.10.071.

Chowdhury, S., Champagne, P., McLellan, P.J., 2008. Factors influencing formation of trihalomethanes in drinking water: results from multivariate statistical investigation of the Ontario Drinking Water Surveillance Program database. Water Qual. Res. J. 43, 189–199. https://doi.org/10.2166/wqrj.2008.022.

Constances [WWW Document], 2023. URL doi.org/10.13143/inserm_constances (accessed 3.21.24).

Corso, M., Galey, C., Beaudeau, P., 2017. Évaluation quantitative de l'impact sanitaire des sous-produits de chloration dans l'eau destinée à la consommation humaine en France. Santé Publique France 44.

Corso, M., Galey, C., Seux, R., Beaudeau, P., 2018. An assessment of current and past concentrations of trihalomethanes in drinking water throughout France. Int. J. Environ. Res. Publ. Health 15, 1669. https://doi.org/10.3390/ijerph15081669.

Costet, N., Garlantézec, R., Monfort, C., Rouget, F., Gagnière, B., Chevrier, C., Cordier, S., 2012. Environmental and urinary markers of prenatal exposure to drinking water disinfection by-products, fetal growth, and duration of gestation in the PELAGIE birth cohort (Brittany, France, 2002–2006). Am. J. Epidemiol. 175, 263–275. https://doi.org/10.1093/aje/kwr419.

Costet, N., Villanueva, C.M., Jaakkola, J.J.K., Kogevinas, M., Cantor, K.P., King, W.D., Lynch, C.F., Nieuwenhuijsen, M.J., Cordier, S., 2011. Water disinfection by-products and bladder cancer: is there a European specificity? A pooled and meta-analysis of European case–control studies. Occup. Environ. Med. 68, 379–385. https://doi.org/10.1136/oem.2010.062703.

Elwood, J.M., van der Werf, B., 2022. Nitrates in drinking water and cancers of the colon and rectum: a meta-analysis of epidemiological studies. Cancer Epidemiol 78, 102148. https://doi.org/10.1016/j.canep.2022.102148.

Espejo-Herrera, N., Gracia-Lavedan, E., Boldo, E., Aragonés, N., Pérez-Gómez, B., Pollán, M., Molina, A.J., Fernández, T., Martín, V., La Vecchia, C., 2016a. Colorectal cancer risk and nitrate exposure through drinking water and diet. Int. J. Cancer 139, 334–346. https://doi.org/10.1002/ijc.30083.

Espejo-Herrera, N., Gracia-Lavedan, E., Pollan, M., Aragonés, N., Boldo, E., Perez-Gomez, B., Altzibar, J.M., Amiano, P., Zabala, A.J., Ardanaz, E., 2016b. Ingested nitrate and breast cancer in the Spanish multicase-control study on cancer (MCC-Spain). Environ. Health Perspect. 124, 1042–1049. https://doi.org/10.1289/ehp.1510334.

European Environment Agency (EEA), 2023. Groundwater nitrate. Available online: https://www.eea.europa.eu/en/analysis/indicators/nitrate-in-groundwater-8th-eap?activeAccordion=ecdb3bcf-bbe9-4978-b5cf-0b136399d9f8. (Accessed 10 June 2024).

Evlampidou, I., Font-Ribera, L., Rojas-Rueda, D., Gracia-Lavedan, E., Costet, N., Pearce, N., Vineis, P., Jaakkola, J.J.K., Delloye, F., Makris, K.C., Stephanou, E.G., Kargaki, S., Kozisek, F., Sigsgaard, T., Hansen, B., Schullehner, J., Nahkur, R., Galey, C., Zwiener, C., Vargha, M., Righi, E., Aggazzotti, G., Kalnina, G., Grazuleviciene, R., Polanska, K., Gubkova, D., Bitenc, K., Goslan, E.H., Kogevinas, M., Villanueva, C.M., 2020. Trihalomethanes in drinking water and bladder cancer burden in the European union. Environ. Health Perspect. 128, 1. https://doi.org/10.1289/EHP4495.

Font-Ribera, L., Gràcia-Lavedan, E., Aragonés, N., Pérez-Gómez, B., Pollán, M., Amiano, P., Jiménez-Zabala, A., Castaño-Vinyals, G., Roca-Barceló, A., Ardanaz, E., Burgui, R., Molina, A.J., Fernández-Villa, T., Gómez-Acebo, I., Dierssen-Sotos, T., Moreno, V., Fernandez-Tardon, G., Peiró, R., Kogevinas, M., Villanueva, C.M., 2018. Long-term exposure to trihalomethanes in drinking water and breast cancer in the Spanish multicase-control study on cancer (MCC-Spain). Environ. Int. 112, 227–234. https://doi.org/10.1016/j.envint.2017.12.031.

Helte, E., Säve-Söderbergh, M., Larsson, S.C., Martling, A., Åkesson, A., 2023. Disinfection by-products in drinking water and risk of colorectal cancer: a population-based cohort study. JNCI J. Natl. Cancer Inst. djad145. https://doi.org/10.1093/jnci/djad145.

Hong, H., Zhang, Z., Guo, A., Shen, L., Sun, H., Liang, Y., Wu, F., Lin, H., 2020. Radial basis function artificial neural network (RBF ANN) as well as the hybrid method of RBF ANN and grey relational analysis able to well predict trihalomethanes levels in tap water. J. Hydrol. 591, 125574 https://doi.org/10.1016/j.jhydrol.2020.125574.

Inoue-Choi, M., Ward, M.H., Cerhan, J.R., Weyer, P.J., Anderson, K.E., Robien, K., 2012. Interaction of nitrate and folate on the risk of breast cancer among postmenopausal women. Nutr. Cancer 64, 685–694. https://doi.org/10.1080/01635581.2012.687427.

International Agency for Research on Cancer, 2012. Some chemicals present in industrial and consumer products, food and drinking-water. IARC Monogr. Eval. Carcinog. Risks Hum. 101.

International Agency for Research on Cancer, 2010. Ingested nitrate and nitrite, and cyanobacterial peptide toxins. IARC Monogr. Eval. Carcinog. Risks Hum. 94.

Kinani, A., Sa Lhi, H., Bouchonnet, S., Kinani, S., 2018. Determination of adsorbable organic halogens in surface water samples by combustion-microcoulometry versus combustion-ion chromatography titration. J. Chromatogr. A 1539, 41–52. https://doi.org/10.1016/j.chroma.2018.01.045.

Koivusalo, M., Pukkala, E., Vartiainen, T., Jaakkola, J.J.K., Hakulinen, T., 1997. Drinking water chlorination and cancer–a historical cohort study in Finland. Cancer Causes Control 8, 192–200. https://doi.org/10.1023/a:1018420229802.

Menard, C., Heraud, F., Volatier, J.-L., Leblanc, J.-C., 2008. Assessment of dietary exposure of nitrate and nitrite in France. Food Addit. Contam. 25, 971–988. https://doi.org/10.1080/02652030801946561.

Mouly, D., Joulin, E., Rosin, C., Beaudeau, P., Zeghnoun, A., Olszewski-Ortar, A., Munoz, J.-F., 2009. Les sous-produits de chloration dans l'eau destinée à la consommation humaine en France. Campagnes d'analyses dans quatre systèmes de distribution d'eau et modélisation de l'évolution des trihalométhanes. Saint-Maurice : Institut de veille sanitaire 73.

Rahman, MdB., Driscoll, T., Cowie, C., Armstrong, B.K., 2010. Disinfection by-products in drinking water and colorectal cancer: a meta-analysis. Int. J. Epidemiol. 39, 733–745. https://doi.org/10.1093/ije/dyp371.

Rossman, L., Clark, R., Grayman, W., 1994. Modeling chlorine residuals in drinking-water distribution systems. J. Environ. Eng. 120, 803–820. https://doi.org/10.1061/(ASCE)0733-9372(1994)120:4(803.

Rossman, L.A., Brown, R.A., Singer, P.C., Nuckols, J.R., 2001. DBP formation kinetics in a simulated distribution system. Water Res. 35, 3483–3489. https://doi.org/10.1016/S0043-1354(01)00059-8.

Schullehner, J., Hansen, B., 2014. Nitrate exposure from drinking water in Denmark over the last 35 years. Environ. Res. Lett. 9, 95001 https://doi.org/10.1088/1748-9326/9/9/095001.

Schullehner, J., Hansen, B., Thygesen, M., Pedersen, C.B., Sigsgaard, T., 2018. Nitrate in drinking water and colorectal cancer risk: a nationwide population-based cohort study. Int. J. Cancer 143, 73–79. https://doi.org/10.1002/ijc.31306.

The European Parliament and the Council of the European Union, 2020. Directive (EU) 2020/2184 of the European parliament and of the Council of 16 december 2020 on the quality of water intended for human consumption. Off. J. Eur. Union 63, 1–62.

Tiouiouine, A., Jabrane, M., Kacimi, I., Morarech, M., Bouramtane, T., Bahaj, T., Yameogo, S., Rezende-Filho, A.T., Dassonville, F., Moulin, M., 2020. Determining the relevant scale to analyze the quality of regional groundwater resources while combining groundwater bodies, physicochemical and biological databases in southeastern France. Water 12, 3476. https://doi.org/10.3390/w12123476.

Tsubura, A., Lai, Y.-C., Miki, H., Sasaki, T., Uehara, N., Yuri, T., Yoshizawa, K., 2011. Animal models of N-methyl-N-nitrosourea-induced mammary cancer and retinal degeneration with special emphasis on therapeutic trials. In Vivo 25, 11–22.

Uyak, V., Toroz, I., Meriç, S., 2005. Monitoring and modeling of trihalomethanes (THMs) for a water treatment plant in Istanbul. Desalination. Seminar in Environmental Science and Technology: Evaluation of Alternative Water Treatment Systems for Obtaining Safe Water 176, 91–101. https://doi.org/10.1016/j.desal.2004.10.023.

Villanueva, C.M., Cantor, K.P., Cordier, S., Jaakkola, J.J.K., King, W.D., Lynch, C.F., Porru, S., Kogevinas, M., 2004. Disinfection byproducts and bladder cancer: a pooled analysis. Epidemiology 357–367. https://doi.org/10.1097/01.ede.0000121380.02594.fc.

Villanueva, C.M., Gracia-Lavedan, E., Bosetti, C., Righi, E., Molina, A.J., Martín, V., Boldo, E., Aragonés, N., Perez-Gomez, B., Pollan, M., 2017. Colorectal cancer and long-term exposure to trihalomethanes in drinking water: a multicenter case–control study in Spain and Italy. Environ. Health Perspect. 125, 56–65. https://doi.org/10.1289/EHP155.

Villanueva, C.M., Kogevinas, M., Cordier, S., Templeton, M.R., Vermeulen, R., Nuckols, J.R., Nieuwenhuijsen, M.J., Levallois, P., 2014. Assessing exposure and health consequences of chemicals in drinking water: current state of knowledge and research needs. Environ. Health Perspect. 122, 213–221. https://doi.org/10.1289/ehp.1206229.

Ward, M.H., Jones, R.R., Brender, J.D., de Kok, T.M., Weyer, P.J., Nolan, B.T., Villanueva, C.M., van Breda, S.G., 2018. Drinking water nitrate and human health: an updated review. Int. J. Environ. Res. Publ. Health 15 (7), 1557.

World Health Organization, 2022. Guidelines for drinking-water quality. In: Fourth Edition Incorporating the First and Second Addenda.

Xu, Z., Shen, J., Qu, Y., Chen, H., Zhou, X., Hong, H., Sun, H., Lin, H., Deng, W., Wu, F., 2022. Using simple and easy water quality parameters to predict trihalomethane occurrence in tap water. Chemosphere 286, 131586. https://doi.org/10.1016/j.chemosphere.2021.131586.

Zins, M., Goldberg, M., Team, C., 2015. The French CONSTANCES population-based cohort: design, inclusion and follow-up. Eur. J. Epidemiol. 30, 1317–1328.