Supplementary material S2: Material and Methods of Medical Research

To study the impact of different types of water sources and water treatment on the risk of arterial hypertension, we collected data from fieldwork based on medical examinations and surveys in the regions between the rivers of Ob, Nadym, Taz, and Yenisey, in 2012, 2014–2019, and 2022. The fieldwork was conducted by researchers of the Sechenov University, the YNAO Arctic Scientific Research Centre, and the Northern Arctic Federal University. A total of 552 local rural residents of the Arctic zone of Western Siberia (138 men; 414 women) participated in the study. The average age was 43.0 [SD: 36.0–55.0]. 218 patients were diagnosed with arterial hypertension. Patients aged from 40 to 69 years old were invited to participate in the study. This age class was chosen due to high potential risk of arterial hypertension in this age group [130]. Other inclusion criteria for the respondents were as follows: a) have resided in the rural areas (settlements and nomadic camps in the tundra) of the Arctic zone of Western Siberia for over ten years; and b) have consumed drinking water from what they identified as a dominant water source at least five times a week for over ten years. Exclusion criteria for the participants were: i) pregnant women; ii) patients who have other chronic diseases in the acute stage such as diseases of the kidneys, central nervous system, and endocrine diseases that affect the water-salt metabolism, iii) patients with anomalies in the development of the kidney, renal artery, aorta and heart; iv) patients taking medications that affect water-salt metabolism; v) those who have had contact with heavy metals due to their professional duties. Data collection was performed in the Russian language and participants filled out a confidential survey.

The survey sequence was as follows: during an expedition between 2012 and 2022, the participants were asked to fill in the questionnaire while undergoing the medical examination. The surveys were taken at health care institutions which were municipal hospitals and paramedic "feldsher"-midwife medical stations in remote settlements to study the impact of water consumption (water sources and water preparation) on human health (primarily on arterial hypertension). Questionnaires included questions for collecting information about sociodemographic factors (age, sex, place of residence, housing conditions) and medical status (chronic diseases, blood pressure, water consumption). The water consumption data was collected according to the following survey questions: "What main water source do you use (a river, a lake, ice, snow, water supply system) for daily consumption?", "How do you mainly prepare water for daily consumption (drink fresh water, settle, filter, boil, freeze water)?" Medical data collection was undertaken by doctors who had been trained in the study procedures. Blood pressure was measured three times, and the presence of hypertension was established in accordance with the recommendations for hypertension [131,132].

The respondents received information about the programme, both verbally and in writing. The written consent form stated that participation was voluntary and that their confidentiality was assured. Participants' personal data and their answers were anonymised, numbered, and entered into de-identified databases.

Statistical Analysis

Based on the survey and medical examinations, a model was developed of the early formation of health changes due to the consumption of water from different water sources. Nonlinear logit regression with stepwise inclusion of variables using the maximum likelihood method was used to build risk models for the development of arterial hypertension. Due to the fact that the dependent variable (presence or absence of arterial hypertension) is dichotomous (binary), we chose a logistic regression [133]. The probability (p) of the onset of arterial hypertension was calculated by the formula (1) [134]:

$$p = \frac{1}{(1 + e^{-z})}, \tag{1}$$

e – a constant (Euler's number) approximately equal to 2.71828,

z – a parameter of the model (formula 2)

$$z = b_1 \times x_1 + b_2 \times x_2 + \ldots + b_n \times x_n + b_0, \tag{2}$$

x_i – a value of independent variables,

b_i – a regression coefficient of logistic regression,

b_0 – a constant (the value of the function under study at zero values of x, when working with categorical independent variables – at their reference values [135].

If the probability of p is less than 0.5, then we can assume that an event (the onset of hypertension) will occur; otherwise, the absence of an event is assumed [134]. A quantitative assessment of the risk of developing hypertension was calculated as the odds ratio in the presence or absence of the studied disease (formula 3):

$$OR = \frac{a \times d}{b \times c}, \tag{3}$$

OR – an odds ratio calculated from the results of the study,
a – number of cases of the presence of the disease in a group of patients,
b – number of cases of the absence of the disease in a group of patients,
c – number of cases of the presence of the disease in a healthy group,
d – number of cases of the absence of the disease in a healthy group [136].

The Chi^2 test was used to assess the significance of the risk of arterial hypertension, and a 95% confidence interval was calculated for the odds ratio.

Statistical analyses were performed using Microsoft Excel 2016 and Statistica for Windows, v. 8.0 (StatSoft Inc., Oklahoma, USA). Significant differences were defined at a p-value < 0.05.

Validity statistics of the model are presented in Table S2.

Table S2. Validity statistics: the impact of the type of water sources on the risk of arterial hypertension.

Odds ratio: 4.6354 Perc. Correct: 69.57%

	Pred. – 0.000000	**Pred. – 1.000000**	**Percent – Correct**
0.000000	272	62	81.43713
1.000000	106	112	51.37615

A quantitative assessment of the risk of developing hypertension was calculated as the odds ratio in the presence or absence of the studied disease (3):

$$OR = \frac{272 \times 112}{106 \times 62} = 4.6354.$$

The overall accuracy of the model is 69.57%, sensitivity 51.37615, specificity 81.43713; odds ratio 4.6354; Chi^2= 96,359, p is less than 0.00001.