Supplementary figures and tables: Regularized Regression Analysis for the Prediction of Virus Inactivation Efficiency by Chloramine Disinfection

Syun-suke Kadoya[a], Osamu Nishimura[a], Hiroyuki Kato[b], Daisuke Sano[a,c*]

[a]Department of Civil and Environmental Engineering, Graduate School of Engineering, Tohoku University, Aoba 6-6-06, Aramaki, Aoba-ku, Sendai, Miyagi 980-8579, Japan

[b]New Industry Creation Hatchery Center, Tohoku University, Sendai, Miyagi, Japan.

[c]Department of Frontier Sciences for Advanced Environment, Graduate School of Environmental Studies, Tohoku University, Aoba 6-6-06, Aramaki, Aoba-ku, Sendai, Miyagi 980-8579, Japan

*Corresponding author: Daisuke Sano, Ph.D., Graduate School of Environmental Studies, Tohoku University, Aoba 6-6-06, Aramaki, Aoba-ku, Sendai, Miyagi 980-8579, Japan

Email: daisuke.sano.e1@tohoku.ac.jp

Tell: +81-11-795-7481

Table S1. The number of strains and the type of strain for each virus species

	N-Strain	*Strain types*
Norovirus	1	Murine norovirus
Adenovirus	4	type 2, 3, 5, 40
Poliovirus	3	type 1, 2, 3
Coxsackievirus	2	B3, 5
Echovirus	2	type1, 11

Table S2. Akaike Information Criterion for deciding the appropriate probability distributions of virus log reduction values (LRVs)

	Normal	*Log Normal*	*Gamma*	*Weibull*	*Exponential*
Norovirus	261	316	288	270	321
Adenovirus	774	*No fit*	*No fit*	*No fit*	816
Poliovirus	211	188	186	187	191
Coxsackievirus	146	158	141	139	142
Echovirus	118	112	105	108	107

Table S3. Variance inflation factors (I: initial concentration of chloramine, k: decay constant, t: contact time, C: Ct-value, p: pH, T: temperature, E: electric conductivity, U: turbidity, W: water type)

	I	***k***	***t***	***C***	***p***	***T***	***E***	***U***	***W***
Norovirus	8.7	8.7	1.7	1.7	1.4	2.5	1.7	2.6	2.6
Adenovirus	1.9	3.1	2.1	3.0	1.6	1.9	3.1	3.0	4.5
Poliovirus	1.9	1.2	1.3	1.5	1.1	1.6	-	-	-
Coxsackievirus	2.5	-	1.2	1.8	1.8	-	2.1	2.1	-
Echovirus	-	-	1.3	1.4	1.3	-	1.6	1.7	-

Table S4. Mean squared errors of training and test datasets in Trial 2

		Lasso		Elastic net		Ridge	
		train	**test**	**train**	**test**	**train**	**test**
Norovirus	linear	0.66	8.85	0.67	6.71	0.61	2.41
	interaction	0.51	7.79	0.51	7.93	0.19	21.8
	quadratic	0.48	0.98	0.48	0.98	0.15	6.49
	cubic	0.23	0.69	0.24	0.69	0.03	4.97×10^2
Adenovirus	linear	0.74	0.30	0.73	0.32	0.70	1.14
	interaction	0.68	0.29	0.68	0.32	0.31	4.76×10^2
	quadratic	0.67	0.30	0.67	0.32	0.31	8.14×10^2
	cubic	0.61	0.49	0.60	0.48	0.17	1.16×10^4
Poliovirus	linear	1.39	2.34	1.40	2.22	1.34	2.45
	interaction	1.35	2.42	1.33	2.39	0.60	53.3
	quadratic	1.39	2.29	1.44	2.07	0.74	69.6
	cubic	1.22	2.25	1.23	2.22	0.22	78.7
Coxsackievirus	linear	0.31	1.05	0.30	1.03	0.29	93.4
	interaction	0.20	6.55×10^2	0.19	4.44×10^2	0.05	5.42×10^3
	quadratic	0.20	1.18×10^3	0.19	3.71×10^3	0.04	1.02×10^9
	cubic	0.06	4.16×10^4	0.05	1.12×10^5	0.01	2.39×10^7
Echovirus	linear	0.16	16.5	0.16	25.5	0.15	64
	interaction	0.09	2.56×10^2	0.08	3.36×10^2	0.02	2.02×10^2
	quadratic	0.09	2.79×10^2	0.09	3.52×10^2	0.03	5.98×10^5
	cubic	0.04	4.49×10^2	0.04	5.30×10^2	0.01	1.20×10^6

Figure S1. Nonlinear relationships between log reduction value (LRV) and Ct-value (*r*: correlation coefficient).

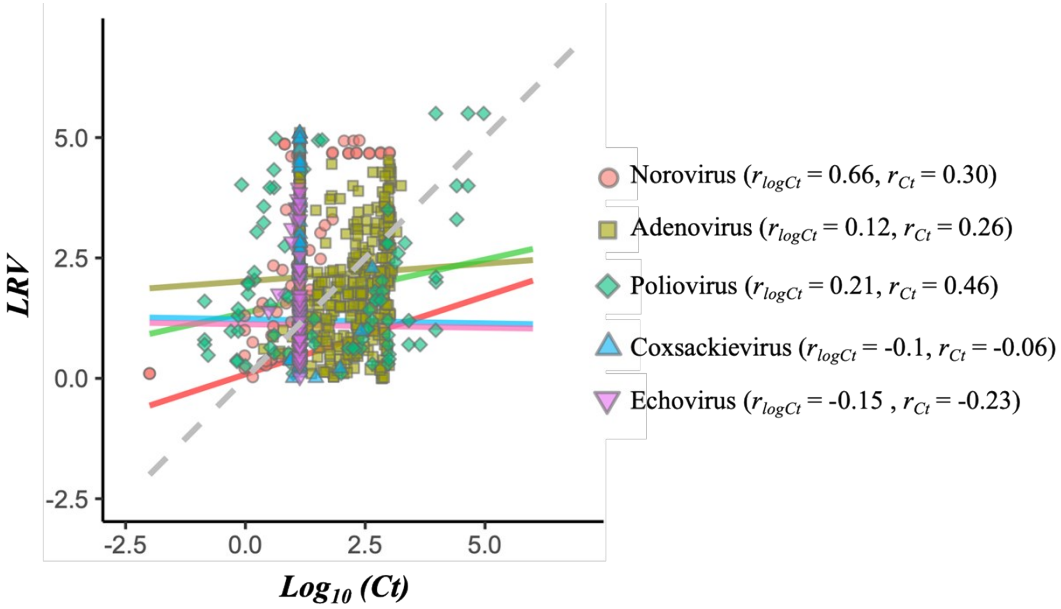

Figure S2. Nonlinear relationships between Ct-value and factors related to Ct-value (initial chloramine concentration, decay constant and contact time). All values were standardized to put them on same panels.

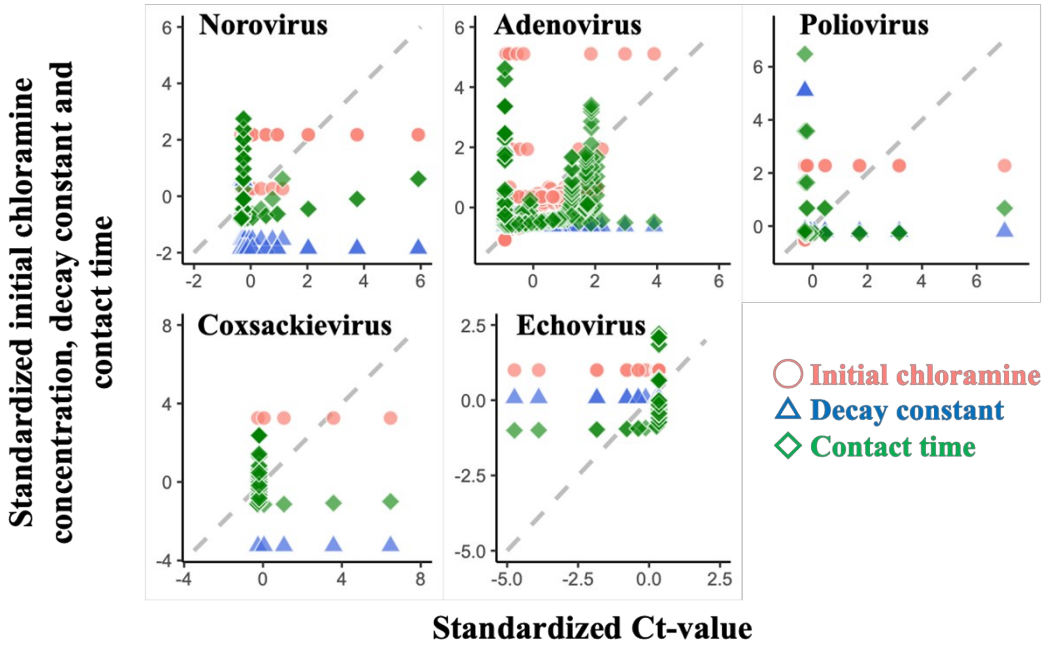

Figure S3. The effect of imputations on the prediction performances (mean: mean imputation, No use: variables having missing values are not employed, zero: zero values are used for missing values, max: zero values are used for missing values).