eAppendix 1: Coefficients and distributions used in the simulation

For the frailty-driven mechanism, we create a full cohort of 100,000 individuals with 10 years of follow-up following eFigure 1A. For every simulated individual, we simulate the time-fixed long-term exposure to $PM_{2.5}$ as a continuous variable (mean=7.4, standard deviation = 2.2) with a lower limit of 1.8 μg/m^3 (background level of $PM_{2.5}$ in Canada).[1] We simulate the time-fixed binary variable *baseline frailty* (p=0.8, with one representing the presence of frailty). We then calculate the *annual probability of death* as $0.006 + 0.0002 \times PM_{2.5} + 0.003 \times$ *baseline frailty* for the first three years and as $0.006 + 0.0002 \times PM_{2.5}$ afterwards. The 0.0002 increase in hazard rate per 1 μg/m^3 increase in $PM_{2.5}$ for the association between $PM_{2.5}$ and mortality is based on Aalen model coefficient from the CCHS cohort. To simulate the variable *time of death*, we simulate a binary variable *event* for each year using the *annual probability of death*, identify the earliest year when the *event* occurred as the *time of death*, and censor the person at the end of 10-year follow-up if no *event* occurred before. The time-fixed *pre-baseline health status* is equal to the reversed first-year *annual probability of death* or $-(0.006 + 0.0002 \times PM_{2.5} + 0.003 \times$ *baseline frailty)*. To simulate an observed cohort considering differential participation, we calculate a *participation score* equal to $0.7 + 10 \times$ *pre-baseline health status* and select those with *participation score* higher than the 30th percentile into the observed cohort, which is similar to the response rate in CCHS Cycle 1.1 and 1.2.[2]

For the geographic factor-driven mechanism, we create a full cohort of 100,000 individuals with 10 years of follow-up following eFigure 1B. In this causal graph, we omitted baseline frailty in Figure 1B for simplicity. For every simulated individual, we simulate the time-fixed binary variable *rural* to represent the geographic factor (p=0.55, with one representing living in a rural area). We simulate the time-fixed long-term exposure to $PM_{2.5}$ as a continuous variable (mean=7.4 – 6 × rural, standard deviation = 2.2) with a lower limit of 1.8 μg/m^3 (background level of $PM_{2.5}$ in Canada).[1] We simulate the *pre-baseline health status* as a binary variable (p= 0.5, with one representing a better than median pre-baseline health status). We then calculate the *annual probability of death* as $0.006 + 0.0002 \times PM_{2.5} - 0.003 \times$ *pre-baseline health status* for the first three years and as $0.006 + 0.0002 \times PM_{2.5}$ afterwards. To simulate the variable *time of death*, we simulate a binary variable *event* for each year using the *annual probability of death*, identify the earliest year when the *event* occurred as the *time of death*, and censor the person at the end of the10-year follow-up if no *event* occurred before. To simulate an observed cohort while considering differential participation, we calculate a *participation score* equal to $0.7 + 0.6 \times$ *rural* $+ 0.45 \times$ *pre-baseline health status* and select those with *participation score* higher than the 30th percentile into the observed cohort.

eAppendix 2: Analyses done for simulated cohorts

For naïve analysis with g-computation, we calculate the difference in probability of survival over time between the 5 $\mu g/m^3$ threshold intervention and natural course with parametric g-computation. Since our simulated dataset has time-fixed exposure, we first run logistic regressions between the probability of death and residential $PM_{2.5}$, indicators of year, and interaction terms between $PM_{2.5}$ and year for the full and observed cohort separately. Next, we made a copy of the observed cohort and updated the exposures based on the intervention (i.e., no change for natural course; change exposure to 5 $\mu g/m^3$ if the exposure is higher than 5 $\mu g/m^3$) so that we could calculate the cumulative survival probabilities standardized to the confounder distributions of the observe cohort under the intervention using coefficients estimated from the full and the observed cohort. Specifically, we predicted the probability of survival for each time point of interest of every subject conditioning on surviving to the start of the time point in the copy using coefficients from the corresponding logistic regressions (estimated using either the full or the observed cohort). We calculated the cumulative survival probabilities at each time point as the cumulative product of conditional probability of survival up to the time point, for every subject. Last, we estimated the average cumulative survival probabilities for each time point as the mean across all subjects. For g-computation with the washout method, we run parametric g-computation after applying the washout method by only using data since the fourth year of follow-up.

For analyses using Cox and Aalen model, we conduct naïve and washout analyses to estimate the association between residential $PM_{2.5}$ and mortality while assuming constant association over time. We only include residential $PM_{2.5}$ as the independent variable in the Aalen and Cox models. We used three, five, and 10 years of follow-up in the naïve analyses, and three, five, and seven years of follow-up in the washout analyses. We used the "survival" package[3] for the Cox model analysis and the "timereg" package[4] for the Aalen model analysis.

References

1. Health Canada. Health Impacts of Air Pollution in Canada 2021 Report. Published March 15, 2021. Accessed June 17, 2021. https://www.canada.ca/en/health-canada/services/publications/healthy-living/2021-health-effects-indoor-air-pollution.html

2. Government of Canada SC. Canadian Community Health Survey (CCHS) Cycle 1.1. Published October 24, 2007. Accessed September 26, 2021. https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=3359

3. Therneau T. survival: A Package for Survival Analysis in R. Published online 2022. https://CRAN.R-project.org/package=survival

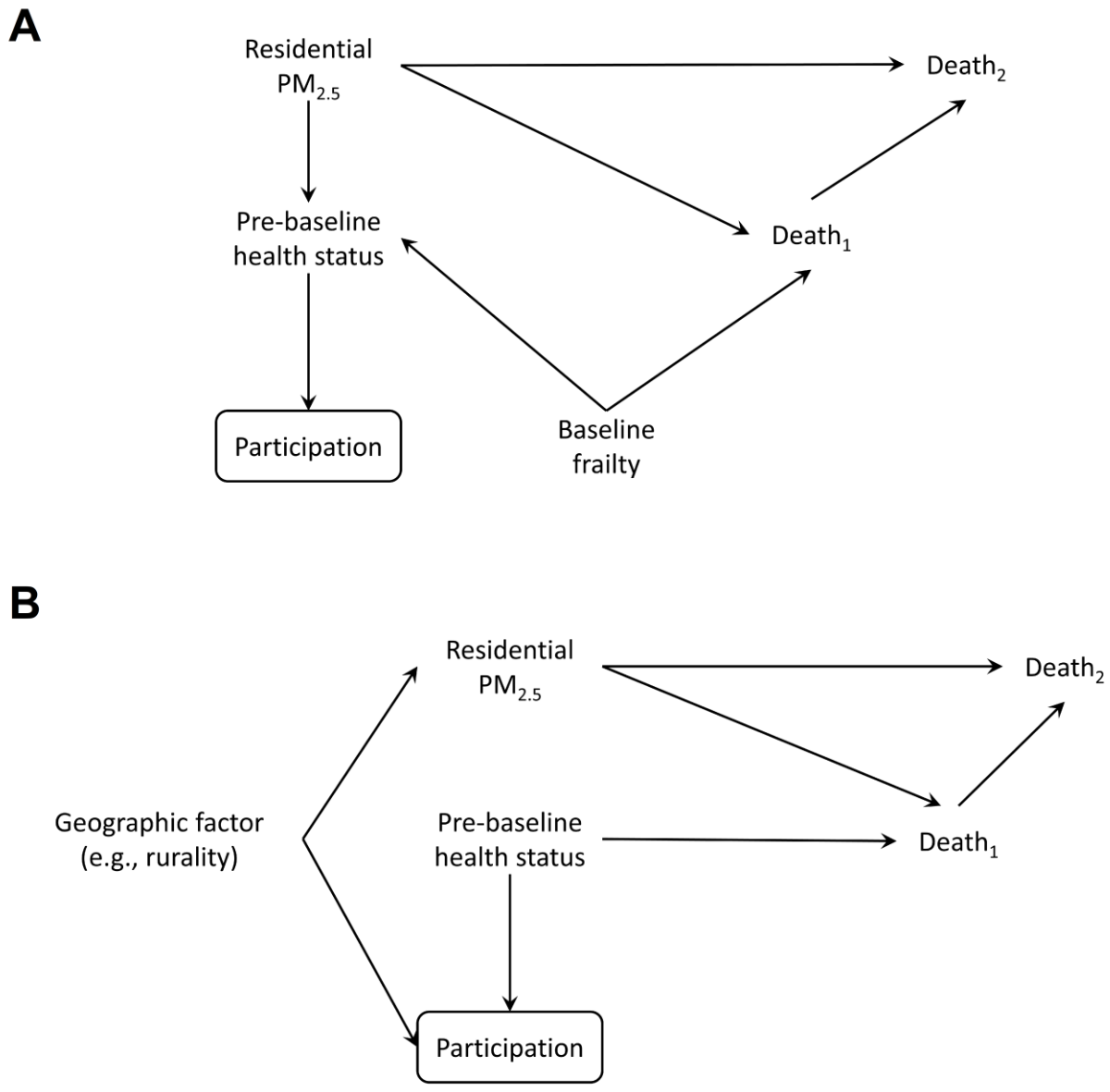
4. Scheike TH, Zhang MJ. Analyzing Competing Risk Data Using the R timereg Package. *Journal of Statistical Software*. 2011;38:1-15. doi:10.18637/jss.v038.i02

eTable 1. Absolute bias (effect estimate in the simulated observed cohort minus the effect estimate in the simulated full cohort) and relative bias (absolute bias divided by the effect estimate in the simulated full cohort) when differential participation exists with and without applying washout in analysis by follow-up time and bias mechanisms (numeric results of Figure 4).

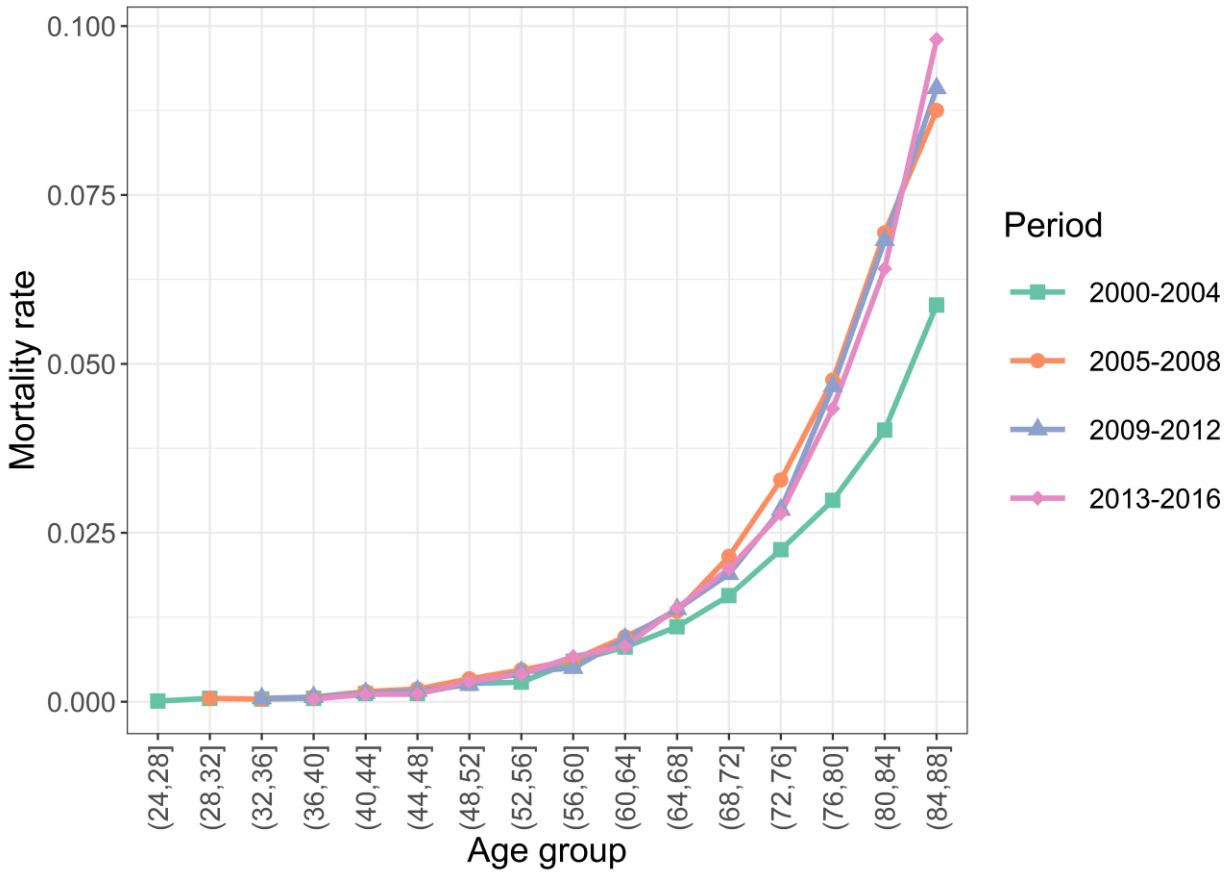
Scenario	Analysis	Follow-up time	Difference in hazard difference per unit change in $PM_{2.5}$ per 1000 persons (95% SI) (Absolute bias-Aalen model)	Percentage change in hazard ratio per unit change in $PM_{2.5}$ (95% SI) (Relative bias-Cox model)
Frailty driven	**Naïve**	**3-year**	-0.16 (-0.36, 0.03)	-2.52 (-4.14, -0.89)
		5-year	-0.14 (-0.27, -0.01)	-1.64 (-3.01, -0.27)
		10-year	-0.07 (-0.16, 0.01)	-0.83 (-1.87, 0.21)
	Washout	**3-year**	0.01 (-0.17, 0.19)	0.14 (-2.05, 2.33)
		5-year	0.00 (-0.15, 0.16)	0.14 (-1.41, 1.70)
		7-year[a]	0.00 (-0.11, 0.12)	0.15 (-1.20, 1.49)
Geographic factor driven	**Naïve**	**3-year**	-0.13 (-0.27, 0.01)	-2.42 (-3.63, -1.21)
		5-year	-0.13 (-0.23, -0.02)	-1.50 (-2.43, -0.58)
		10-year	-0.06 (-0.12, 0.00)	-0.74 (-1.36, -0.12)
	Washout	**3-year**	-0.01 (-0.11, 0.10)	0.02 (-1.05, 1.10)
		5-year	0.00 (-0.08, 0.09)	0.08 (-0.79, 0.94)
		7-year[a]	0.00 (-0.07, 0.08)	0.08 (-0.69, 0.86)

[a]In the simulation, we created a cohort with 10 years of follow-up time thus we only have seven years of follow-up time after dropping the first three years of follow-up in the washout analyses.

eFigure 1. Causal graphs used in simulating two mechanisms in which differential participation could cause a spurious association between residential $PM_{2.5}$ and mortality, based on the CCHS cohort example. $Death_1$ and $Death_2$ represent death during or after first three years of follow-up. A: frailty-driven mechanism; B: geographic factor-driven mechanism.

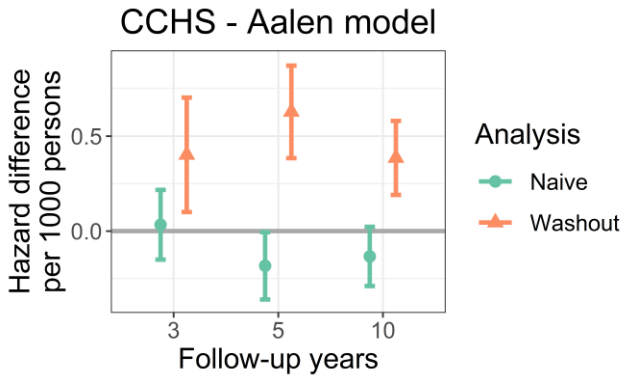


eFigure 2. Age-specific mortality rate in the Canadian Community Health Survey cohort by period.

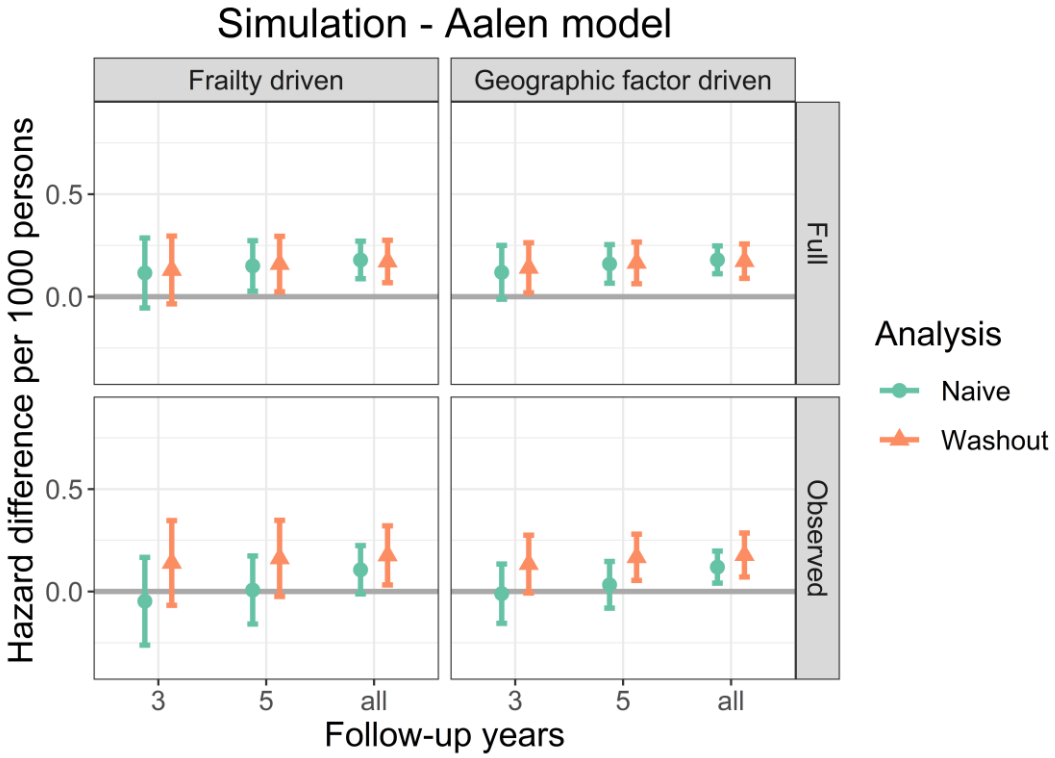


eFigure 3. Effect estimates per unit change in $PM_{2.5}$ when differential participation exists with and without applying washout method by cohort and follow-up time. A: Aalen model results from CCHS cohort (corresponding to observed cohorts in B); B: Aalen model results from simulation cohorts; C: Cox model results from CCHS cohort (corresponding to observed cohorts in D); D: Cox model results from simulation. Note: "all" represents 10-year follow-up in naïve analysis and 7-year follow-up in washout analysis.

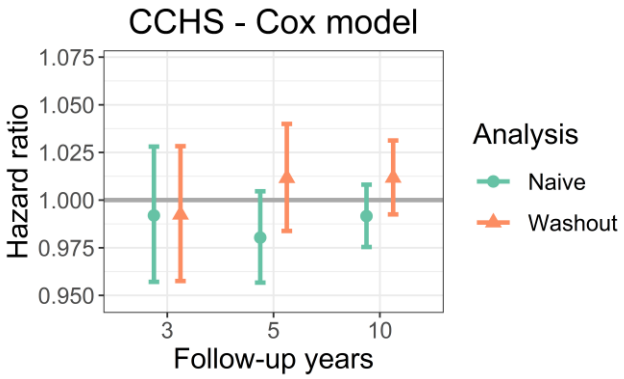
A



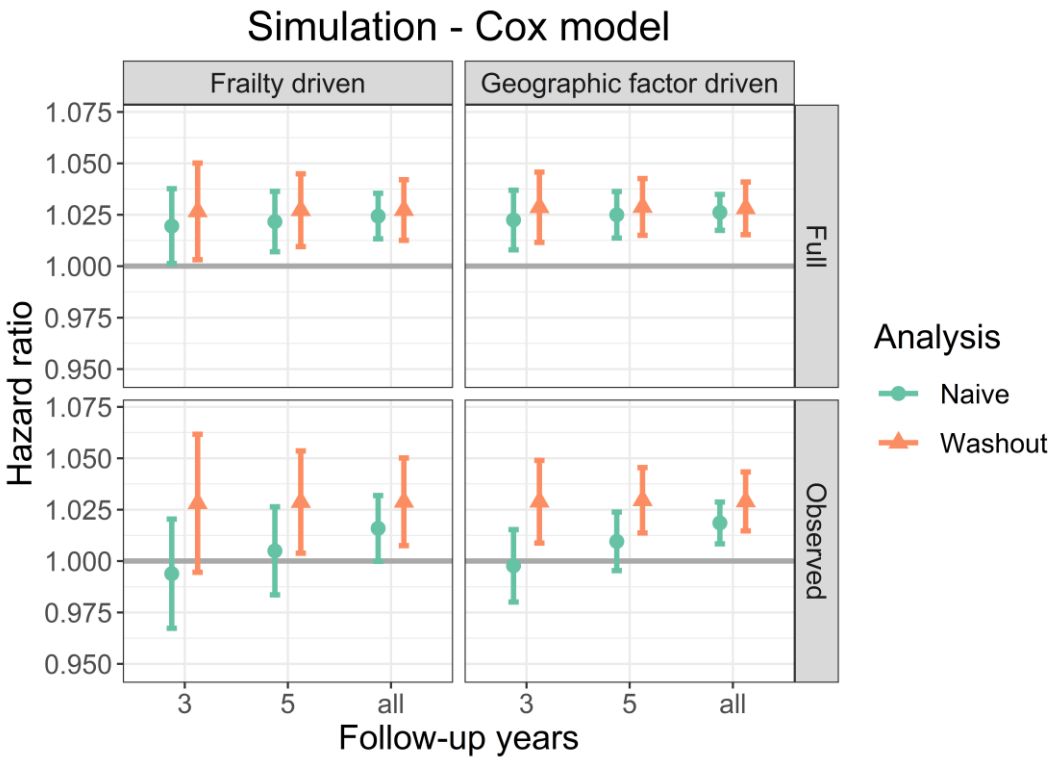
B



C


CCHS - Cox model
Hazard ratio
1.075
1.050
1.025
1.000
0.975
0.950
3
5
10
Follow-up years
Analysis
Naive
Washout

D


Simulation - Cox model
Frailty driven
Geographic factor driven
Full
Observed
Hazard ratio
1.075
1.050
1.025
1.000
0.975
0.950
3
5
all
Follow-up years
Analysis
Naive
Washout