WEB MATERIAL

Causal Inference in Studying the Long-Term Health Effects of Disasters: Challenges and Potential Solutions

Koichiro Shiba, Takuya Kawahara, Jun Aida, Katsunori Kondo, Naoki Kondo, Peter James, Mariana Arcaya, and Ichiro Kawachi

Table of Contents

WEB APPENDIX 1

Steps to Estimate Parametric Survival Curves With Confounding Adjustment via Standardization

Parametric survival curves with confounding adjustment via standardization can be done as follows.

1. Fit a pooled logistic regression to estimate conditional discrete hazards (Step 1 and Step 2 in Figure 1) (1).

   $$logitPr[D_{t+1} = 1|D_t = 0, A, C] = \gamma_{0,t} + \gamma_1 A + \gamma_2 A * t + \gamma_3 A * t^2 + C\gamma'$$

   where D_t is an indicator of event onset at time t, A is the exposure of interest, and C is a vector of covariates that suffices to adjust for all confounding for the association between A and the endpoint D_t.

   The baseline hazard $\gamma_{0,t}$ is a function of time t. For instance, it can be modeled as a quadratic function of time $\gamma_{0,t} = \gamma_0 + \gamma_4 t + \gamma_5 t^2$ or using more flexible specifications such as cubic splines (2). Note that the product terms between the exposure A and time allows hazard "ratios" to vary by time, relaxing the proportional hazards assumption of a Cox model.

2. For each individual, predict conditional discrete hazards $Pr[D_{t+1} = 1|D_t = 0, A, C]$ for each time interval under all possible treatment levels (Step 3 in Figure 1). For instance, when the exposure is binary, simulate trajectories of the discrete hazards for each individual under A=0 and A=1. Use the predicted hazards to obtain corresponding conditional probabilities of survival for each time interval $Pr[D_{t+1} = 0|D_t = 0, A, C] = 1 - Pr[D_{t+1} = 1|D_t = 0, A, C]$. Cumulative probabilities of

survival up to a given time point for each individual can be computed simply by calculating the product of the time-specific survival probabilities.

$$Pr[D_{t^*+1} = 0|A, C] = \prod_{t=0}^{t^*} P\,r[D_{t+1} = 0|D_t = 0, A, C]$$

3. Calculate the mean of the conditional cumulative survival probabilities across individuals for each time point, which effectively estimates marginal cumulative survival probabilities by standardizing the conditional probabilities over the empirical distributions of covariates (Step 4 in Figure 1). Plots of these means over time are the counterfactual survival curves that would have been observed had everyone been exposed to A=a.

One can also plot trajectories of cumulative incidence differences/ratios by comparing two counterfactual cumulative incidence rates under different levels of the exposure over time. Bootstrapping is recommended to compute confidence intervals for the curves (3).

WEB APPENDIX 2

Multivariable-Adjusted Outcome Regression and Inverse Probability Censoring Weighs to Address Selection Bias

1. Multivariable-adjusted outcome regression

The first method is to fit a conventional multivariable adjusted Poisson regression conditional on the exposure A and a vector of covariates C to individuals with observed outcomes (i.e., S=0) (4). One can fit a model such as the one represented by the following equation.

$$logE[Y|A,C,S=0] = \beta_0 + \beta_1 A + C\beta'$$

Under the assumption of no unmeasured common cause for A-Y and S-Y (i.e., conditional exchangeability for treatment and censoring) as well as consistency and positivity, the estimated risk ratio represents the counterfactual estimand of interest conditional on C (see **Web Appendix 4** for a brief proof).

$$exp(\beta_1) = \frac{Pr[Y=1|A=1,C,S=0]}{Pr[Y=1|A=0,C,S=0]} = \frac{Pr[Y^{a=1,s=0}=1|C]}{Pr[Y^{a=0,s=0}=1|C]}$$

2. Inverse probability censoring weights

A second method is to use inverse probability weighting for censoring (IPCWs), $\frac{1}{Pr[S=0|A,C]}$, where the denominator is the probability of no censoring (S=0) conditional on the exposure A and covariates C. IPCWs can be estimated with the following steps.

Step 1. Let D_i be an indicator of death between the disaster onset and outcome assessment for individual i. Use logistic regression to estimate probabilities of not being

censored due to death conditional on A and C. One example of such logistic regression model is the following.

$$logit(\Pr[D_i = 0|A, C]) = \alpha_0 + \alpha_1 A + C\alpha'$$

Step 2. Likewise, estimate probabilities of not being censored due to non-participation among people who were alive at the time of outcome assessment. We again use logistic regression to estimate this probability. Let W_i be an indicator of being alive but non participating in the follow-up survey (withdrawal) for individual i.

$$logit(\Pr[W_i = 0|A, C, D_i = 0]) = \gamma_0 + \gamma_1 A + C\gamma'$$

Step 3. Multiply the two probabilities computed above to estimate probabilities of no censoring $S_i = 0$ conditional on A and C.

$$Pr[S_i = 0|A, C] = Pr[D_i = 0, W_i = 0|A, C] = Pr[W_i = 0|A, C, D_i = 0]Pr[D_i = 0|A, C]$$

Fitting models for $logit(Pr[D_i = 0|A, C])$ and $logit(Pr[W_i = 0|A, C, D_i = 0])$ separately allows differential contributions of covariates to two different censoring mechanisms.

Step 4. Calculate weights as inverse probabilities of no censoring $\frac{1}{Pr[S_i=0|A,C]}$.

The estimated IPCWs are applied to reconstruct the censored individuals by cloning the uncensored people with greater weights (i.e., individuals with low probabilities of no censoring), and creating a pseudo-population in which there is no selective censoring (5). The estimated weights can be multiplied by inverse probability treatment weights (IPTWs), $\frac{1}{Pr[A=a|C]}$, to further adjust for confounding. The final weights can be applied to a Poisson regression for the outcome Y conditional on the exposure A only to estimate a risk ratio.

$$logE[Y|A, S = 0] = \beta_0^* + \beta_1^* A$$

Under conditional exchangeability for both exposure and censoring, consistency, and positivity, the estimated risk ratio represents the counterfactual estimand of interest marginally.

$$exp(\beta_1^*) = \frac{Pr[Y = 1|A = 1, S = 0]}{Pr[Y = 1|A = 0, S = 0]} = \frac{Pr[Y^{a=1,s=0} = 1]}{Pr[Y^{a=0,s=0} = 1]}$$

Confidence intervals are obtained by robust standard errors or bootstrapping (3, 6).

WEB APPENDIX 3

Steps to Estimate Survivor Average Causal Effect

Tchetgen Tchetgen et al (7) proposed the following steps to estimate the survivor average causal effect (SACE).

1. Fit logistic regression models to predict probabilities of no censoring conditional on A and C.

   $$\hat{\pi} = Pr[S = 0|A, C]$$

   Note that this is the denominator of the IPCWs that we discussed above.

2. Fit a Poisson model for the outcome conditional on the exposure A, probabilities of being censored $\hat{Q} = 1 - \hat{\pi}$, and the same set of covariates C as the exposure model using data from people without censoring (S=0).

   $$logE[Y|A, C, S = 0] = \theta_0 + \theta_1 A + \beta_2 \hat{Q} + C\theta'$$

3. The SACE of interest conditional on the covariates C is identified by the regression coefficient of exposure A

   $$exp(\theta_1) = \frac{Pr[Y^{a=1} = 1|S^{a=1} = S^{a=0} = 0, C]}{Pr[Y^{a=0} = 1|S^{a=1} = S^{a=0} = 0, C]}$$

Standard errors and confidence intervals for SACE are obtained via bootstrapping (3). Notably, conditioning on the factor $\hat{Q}$ sufficiently accounts for selection bias due to unmeasured common causes of censoring and outcome (i.e., U) under the assumptions for distributions of U and model specifications listed in Tchetgen Tchetgen et al. (7). A key assumption is the cross-world exchangeability for the S-Y relationship **(**$Y^a \perp\!\!\!\perp S^{1-a}|S^a, A, C, U$**)**, which is implied by the directed acyclic graph in Figure 2C if it was interpreted as non-parametric structural equation models with independent errors. If there is no unmeasured common cause U (i.e., Figure 2B), we

no longer need to condition on $\hat{Q}$ and the standard Poisson regression conditional on A and C is sufficient to identify the same SACE estimand (see Web Appendix 4 for a brief proof).

WEB APPENDIX 4

Identification of the Counterfactual Estimands to Avoid Selection Bias Due to Selective Attrition Using a Conventional Multivariable Regression

Proof 1.

The direct acyclic graph in Figure 2B indicates we have conditional exchangeability for treatment ($Y^{a,s=0} \coprod A|C$) and conditional exchangeability for selection ($Y^{a,s=0} \coprod S|A,C$). Under these assumptions and consistency, we can show that the expected value of the outcome Y conditional on the exposure A=a and a covariate vector C can be interpreted as the mean counterfactual outcome had everyone received A = a and no one been censored conditional on C.

$$\Pr[Y = 1|A = a, C, S = 0] = \Pr[Y^{a,s=0} = 1|A = a, C, S = 0]\ (\because consisency)$$

$$= \Pr[Y^{a,s=0} = 1|A = a, C]\ (\because Y^{a,s=0} \coprod S|A, C)$$

$$= \Pr[Y^{a,s=0} = 1|C] (\because Y^{a,s=0} \coprod A|C)$$

Proof 2.

Under the cross-world exchangeability implied by the directed acyclic graph in Figure 2B ($Y^a \coprod S^{1-a}|S^a, C$) and consistency, we can show that the expected value of the outcome Y conditional on the exposure A=a and a covariate vector C estimated by multivariable adjustment can be interpreted as the mean counterfactual outcome had everyone received A=a among people who would have been uncensored regardless of the levels of the exposure conditional on C estimated by the SACE approach. A short proof is shown below.

1. By consistency, $S^a = S$ and $Y^a = Y$ conditional on $A = a$.

$$\Pr[Y = 1|A = a, C, S = 0] = \Pr[Y^a = 1|A = a, C, S^a = 0] (\because consisency)$$

2. By conditional exchangeability for the exposure, we can drop the conditioning on A=a

$$\Pr[Y^a = 1 | A = a, C, S^a = 0] = \Pr[Y^a = 1 | C, S^a = 0] \; (\because Y^a \perp\!\!\!\perp A | C)$$

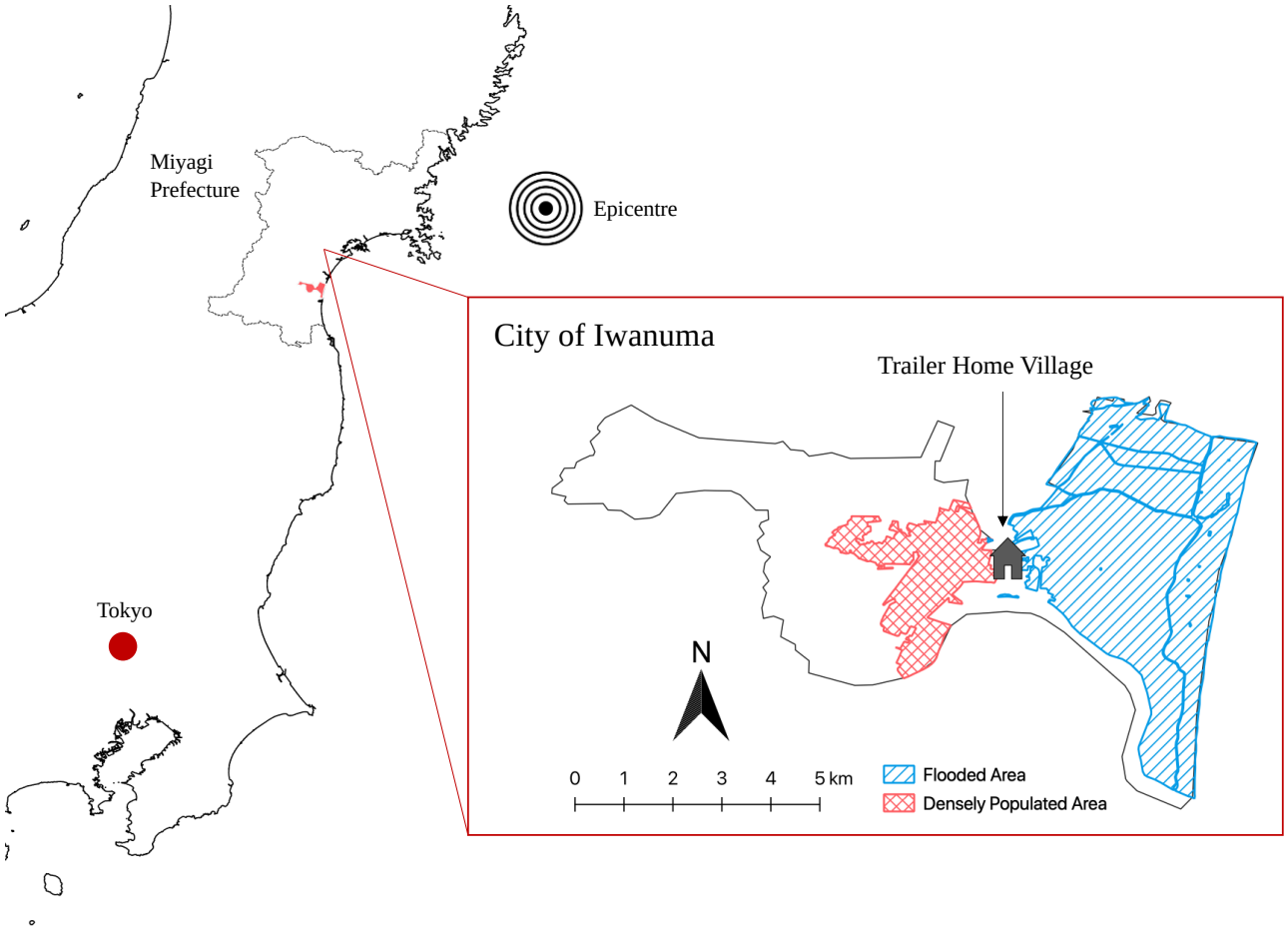
3. By the cross-world exchangeability, we can add conditioning on $S^{1-a} = 0$ conditional on S^a and C.

$$\Pr[Y^a = 1 | C, S^a = 0] = \Pr[Y^a = 1 | C, S^a = S^{1-a} = 0] (\because Y^a \perp\!\!\!\perp S^{1-a} | S^a, C)$$

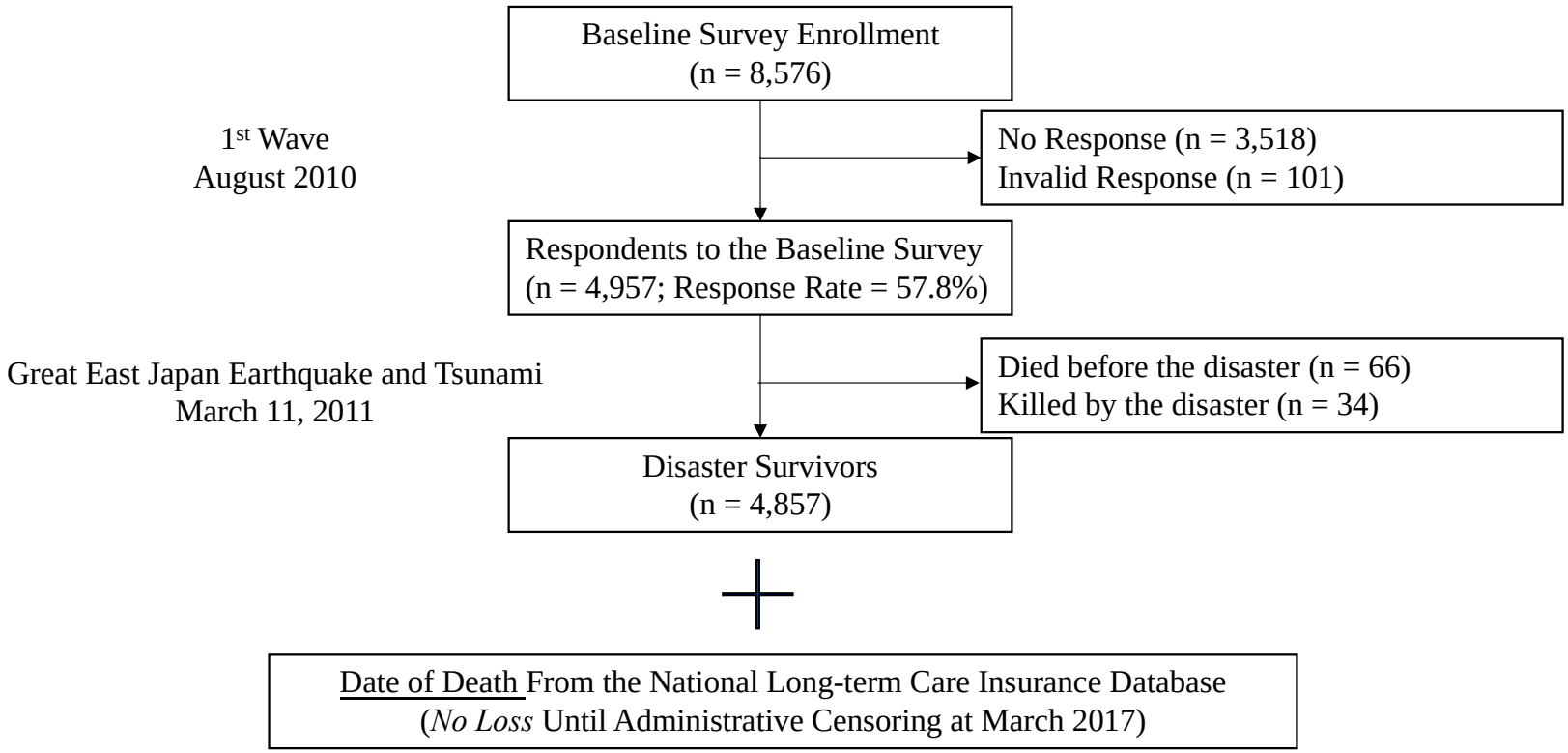
Web Table 1. Trajectories of cumulative incidence difference and cumulative incidence ratio from causal survival analysis

Time	Model 1											
	Cumulative Incidence Difference						Cumulative Incidence Ratio					
	1,000 m–3,000 m (vs >3,000 m)			<1,000 m (vs >3,000 m)			1,000 m–3,000 m (vs >3,000 m)			<1,000 m (vs >3,000 m)		
	Estimate	95% CI		Estimate	95% CI		Estimate	95% CI		Estimate	95% CI	
12 Months	0.015	0.003	0.031	0.034	0.013	0.064	1.65	1.11	2.59	2.50	1.54	3.65
24 Months	0.024	0.004	0.047	0.054	0.025	0.092	1.51	1.08	2.16	2.16	1.51	2.96
36 Months	0.030	0.003	0.062	0.065	0.032	0.112	1.41	1.04	1.93	1.91	1.45	2.48
48 Months	0.034	0.003	0.073	0.071	0.032	0.118	1.34	1.03	1.76	1.72	1.34	2.17
60 Months	0.039	0.005	0.082	0.072	0.027	0.116	1.31	1.04	1.64	1.57	1.22	1.93
72 Months	0.047	0.009	0.093	0.071	0.024	0.123	1.30	1.06	1.59	1.45	1.14	1.79
Time	Model 2											
	Cumulative Incidence Difference						Cumulative Incidence Ratio					
	1,000 m–3,000 m (vs >3,000 m)			<1,000 m (vs >3,000 m)			1,000 m–3,000 m (vs >3,000 m)			<1,000 m (vs >3,000 m)		
	Estimate	95% CI		Estimate	95% CI		Estimate	95% CI		Estimate	95% CI	
12 Months	0.013	0.000	0.029	0.015	-0.000	0.033	1.53	1.01	2.37	1.62	1.00	2.50
24 Months	0.020	0.000	0.045	0.022	0.001	0.050	1.40	1.00	2.00	1.46	1.02	2.04
36 Months	0.024	-0.002	0.056	0.024	0.001	0.057	1.31	0.97	1.76	1.33	1.01	1.79
48 Months	0.027	-0.006	0.066	0.023	-0.006	0.055	1.26	0.95	1.63	1.22	0.94	1.57
60 Months	0.030	-0.003	0.069	0.019	-0.013	0.051	1.23	0.98	1.52	1.14	0.90	1.42
72 Months	0.037	0.002	0.075	0.013	-0.026	0.053	1.23	1.01	1.48	1.08	0.85	1.34

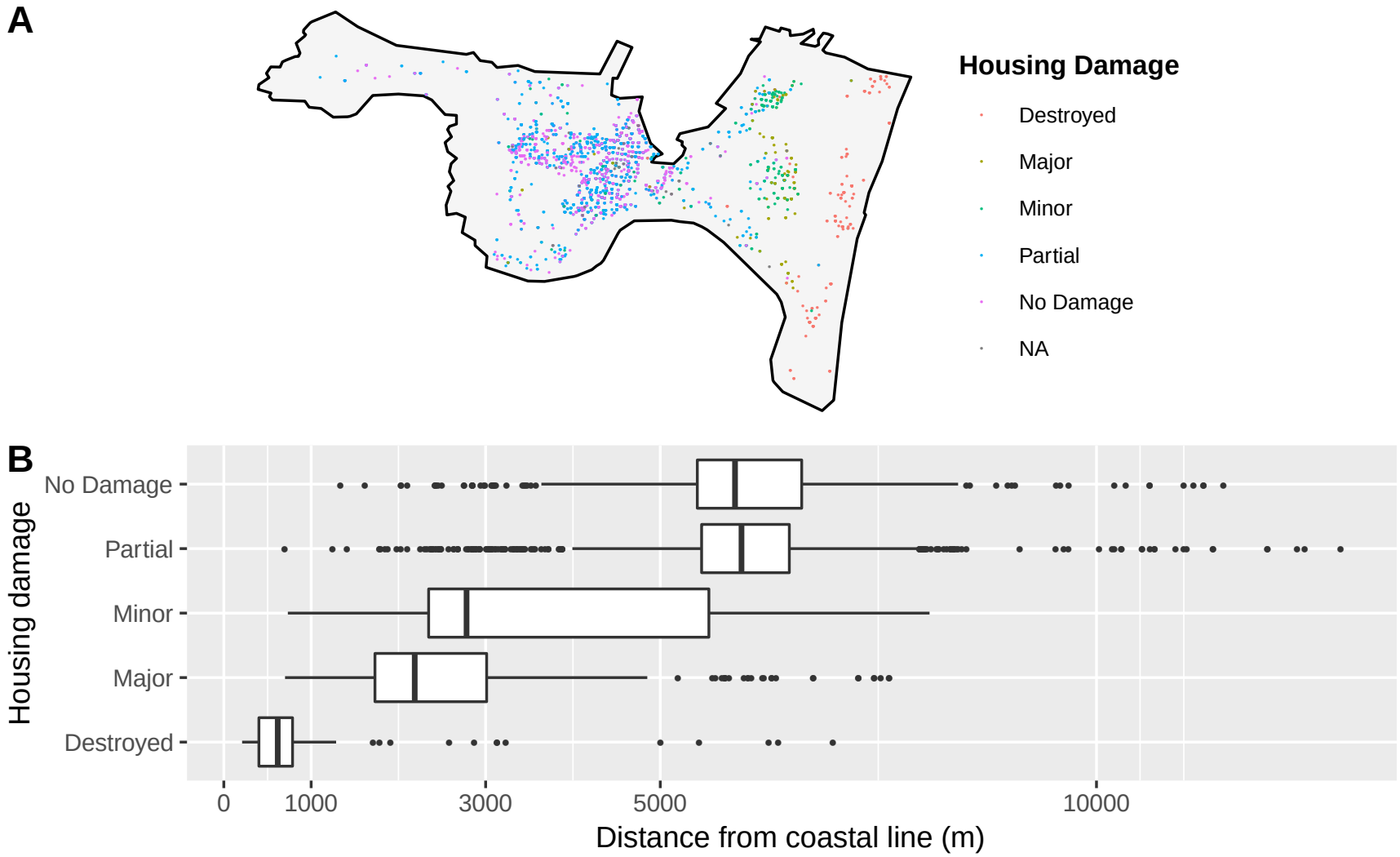
Model 1 examined crude associations between predisaster distance from the coast and mortality. Model 2 was adjusted for gender, age, depressive symptoms, self-rated health, education, household income, current smoking, current alcohol intake, treatment for major diseases including hypertension, stroke, diabetes, and dyslipidemia. Confidence intervals were obtained via bootstrapping with 1000 replications.



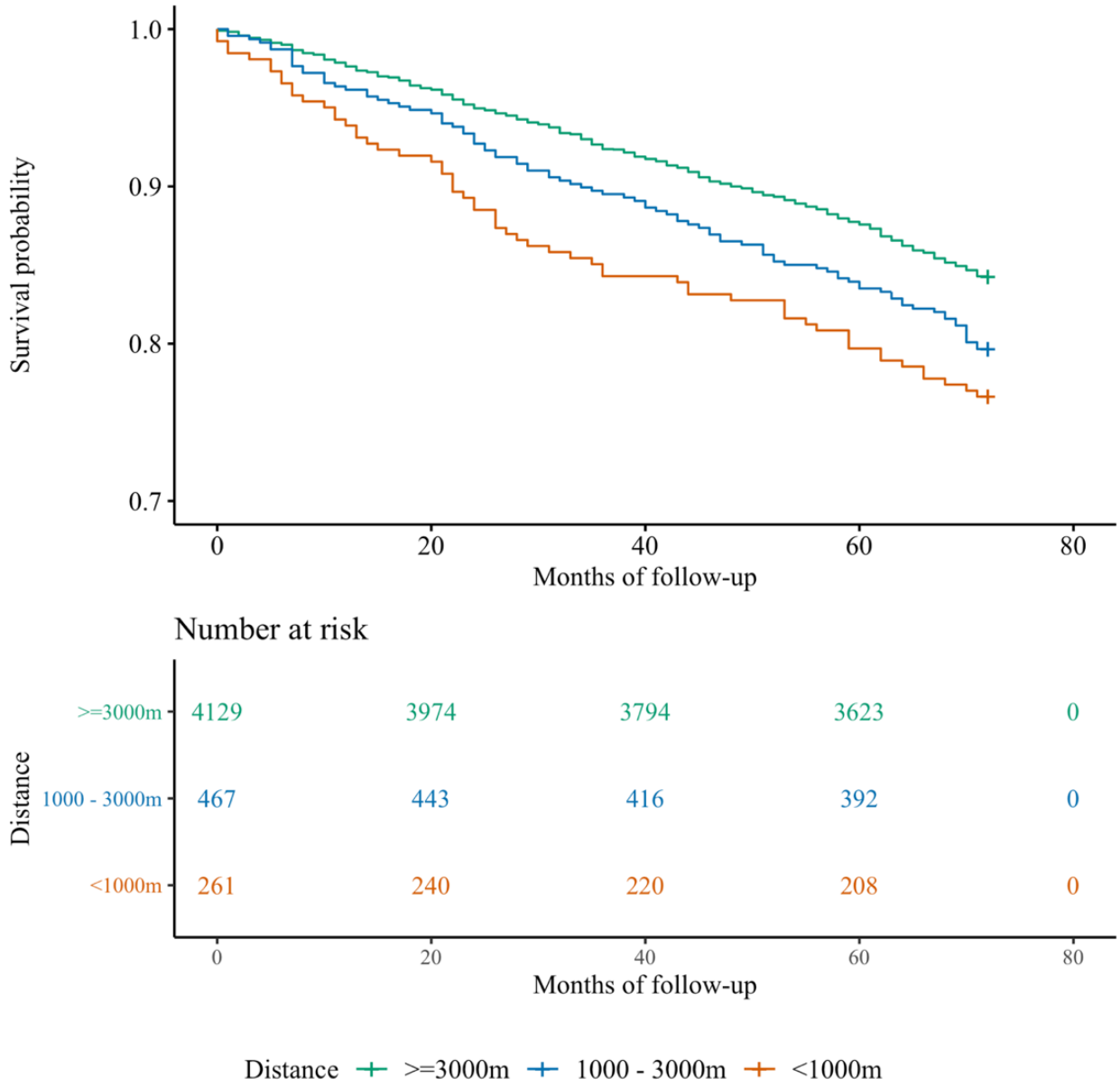
Web Figure 1. Map of Iwanuma City, Miyagi Prefecture, Japan



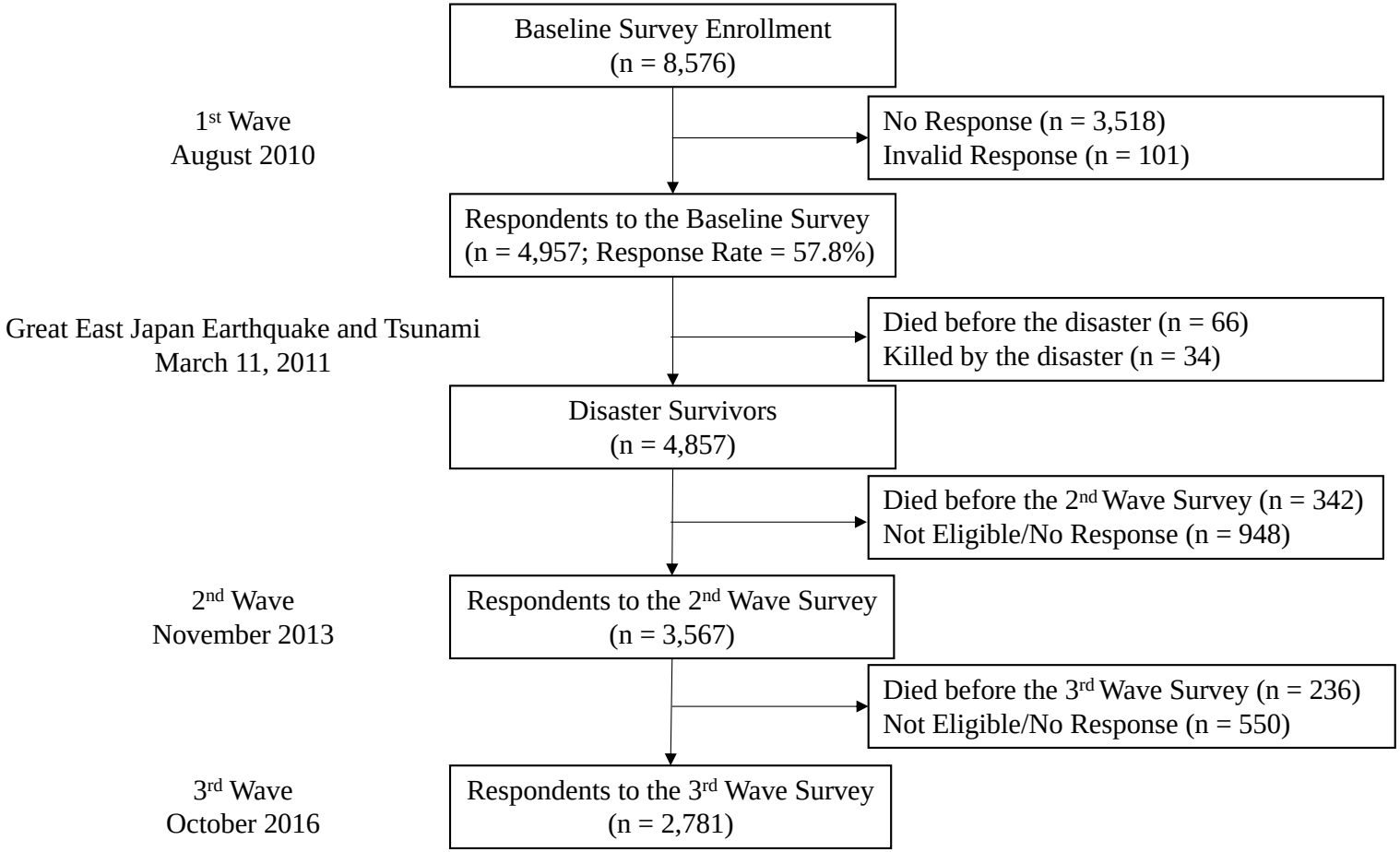
Web Figure 2. Flow of analytical sample selection for demonstration of causal survival analysis.



Web Figure 3. A) Geographic distribution of housing damage in 2013; B) association between pre-earthquake distance from the coast and housing damage.



Web Figure 4. Kaplan-Meier curves for associations between predisaster distance from the coast and mortality.



Web Figure 5. Flow of analytical sample selection for demonstration of selection bias adjustment

Web References

1. D'Agostino RB, Lee M-L, Belanger AJ, et al. Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham Heart Study. *Stat Med.* 1990;9(12):1501–1515.

2. Durrleman S, Simon R. Flexible regression models with cubic splines. *Stat Med*. 1989;8(5):551–561.

3. Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat Sci*. 1986;1(1):54–75.

4. Zou G. A modified Poisson regression approach to prospective studies with binary data. *Am J Epidemiol*. 2004;159(7):702–706.

5. Hernán M, Robins J. IP weighting and marginal structural models. In: *Causal Inference: What If*. Boca Raton, FL: Chapman & Hall/CRC Press; 2020:149–158.

6. Hardin JW. Generalized estimating equations (GEE). In: *Encyclopedia of Statistics in Behavioral Science*. Hoboken, NJ: John Wiley & Sons, Inc.; 2005.

7. Tchetgen Tchetgen EJ, Phiri K, Shapiro R. A simple regression-based approach to account for survival bias in birth outcomes research. *Epidemiology*. 2015;26(4):473–480.