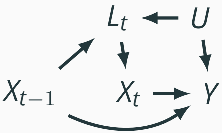
Concepts and Methods for Healthy Worker Survivor Bias

Alex Keil

December 2024

Earl Stadtman Investigator, National Cancer Institute



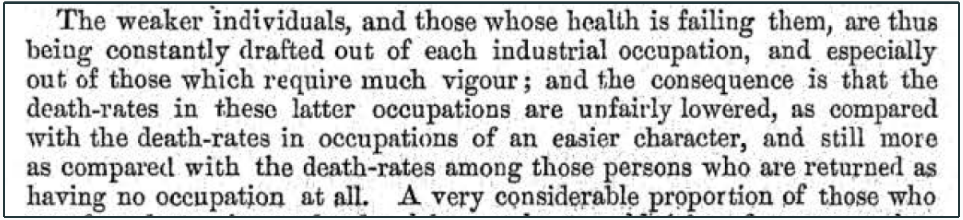
What to expect

- Didactic
- Fast
- Replicable

History: healthy worker survivor effect

Employment and health

Mortality has long been observed to be lower among populations of active workers

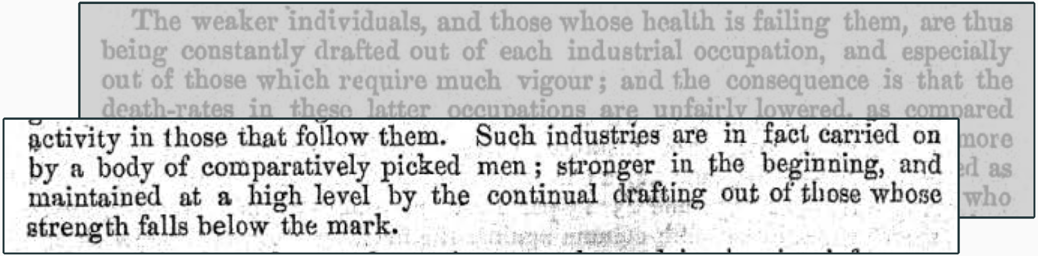


The weaker individuals, and those whose health is failing them, are thus being constantly drafted out of each industrial occupation, and especially out of those which require much vigour; and the consequence is that the death-rates in these latter occupations are unfairly lowered, as compared with the death-rates in occupations of an easier character, and still more as compared with the death-rates among those persons who are returned as having no occupation at all. A very considerable proportion of those who

Ogle (1885) *Supplement to the Forty-Fifth Annual Report of the Registrar-General of Births, Deaths, and Marriages in England*

Employment and health

Mortality has long been observed to be lower among populations of active workers



The weaker individuals, and those whose health is failing them, are thus being constantly drafted out of each industrial occupation, and especially out of those which require much vigour; and the consequence is that the death-rates in these latter occupations are unfairly lowered, as compared

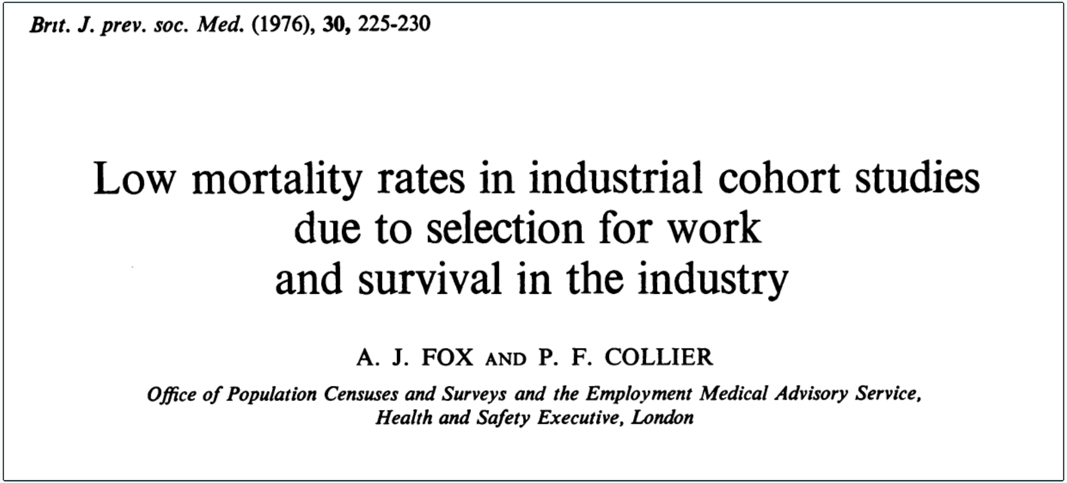
more

d as

who

activity in those that follow them. Such industries are in fact carried on by a body of comparatively picked men; stronger in the beginning, and maintained at a high level by the continual drafting out of those whose strength falls below the mark.

Ogle (1885) *Supplement to the Forty-Fifth Annual Report of the Registrar-General of Births, Deaths, and Marriages in England*

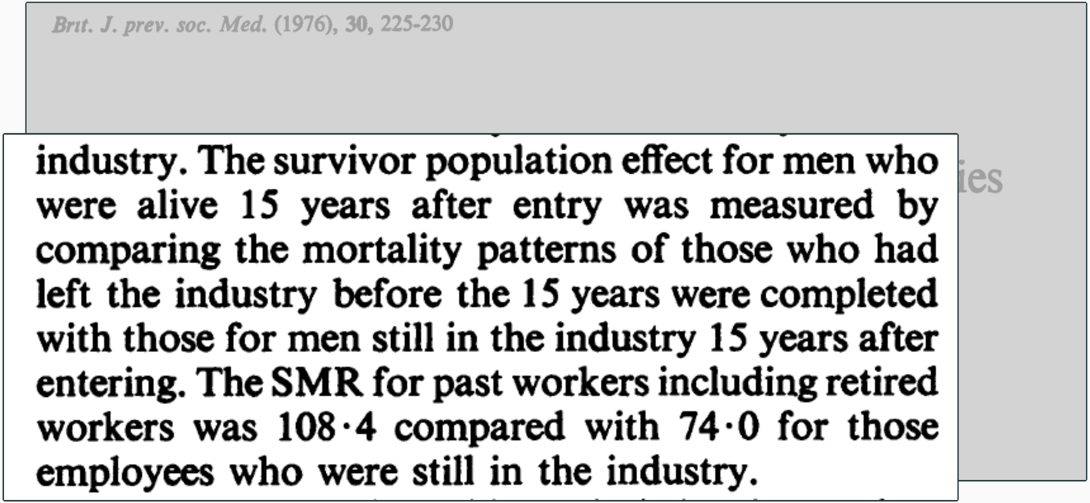


Brıt. J. prev. soc. Med. (1976), **30,** 225-230

Low mortality rates in industrial cohort studies due to selection for work and survival in the industry

A. J. FOX AND P. F. COLLIER

Office of Population Censuses and Surveys and the Employment Medical Advisory Service,
Health and Safety Executive, London

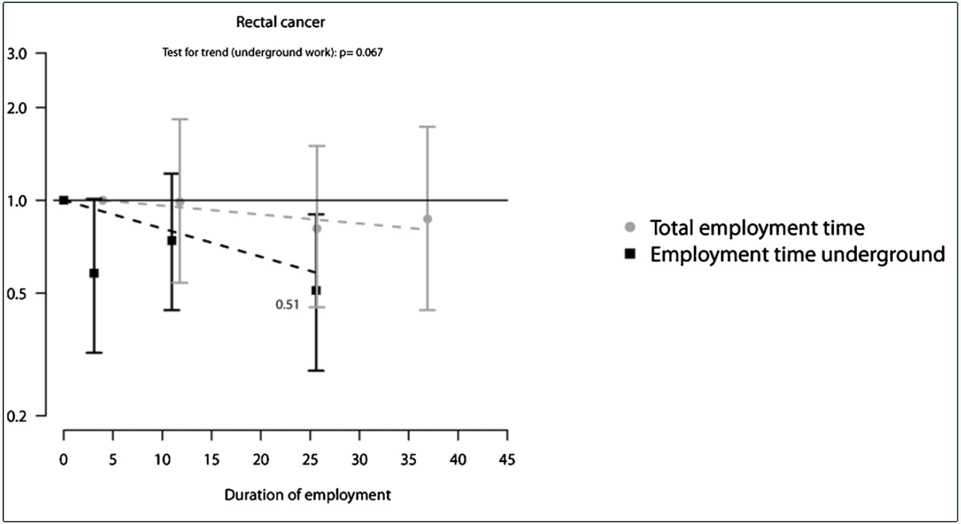


Brıt. J. prev. soc. Med. (1976), 30, 225-230

ies

industry. The survivor population effect for men who were alive 15 years after entry was measured by comparing the mortality patterns of those who had left the industry before the 15 years were completed with those for men still in the industry 15 years after entering. The SMR for past workers including retired workers was 108·4 compared with 74·0 for those employees who were still in the industry.

Longer employment predicts lower risk in 13,623 iron miners



Björ et al. (2013) *Am J Ind Med*

Arsenic and cardiovascular disease among 8,014 copper smelter workers

> *There were* **no significant trends in the relative risks** *with the various measures of cumulative arsenic exposure. However, relative risks for 5–14, 15–24, and 25 years and more after cessation of employment were 0.80, 0.66, and 0.41, respectively, compared with 1–4 years after cessation of employment. These results do not suggest a relation between cardiovascular disease and cumulative arsenic exposure, but do suggest that* **cardiovascular disease "caused the retirement."**

Lubin and Fraumeni Jr (2000) *Am J Epidemiol*

Arsenic and cardiovascular disease among 8,014 copper smelter workers: comment

Lubin and Fraumeni as suggest[ed] that cardiovascular disease "caused the retirement." This is an **excellent description of how the healthy survivor effect operates***...*

Hertz-Picciotto, Hu, and Arrighi (2000) *Am J Epidemiol*

Arsenic and cardiovascular disease in 8,014 copper smelter workers, reanalysis

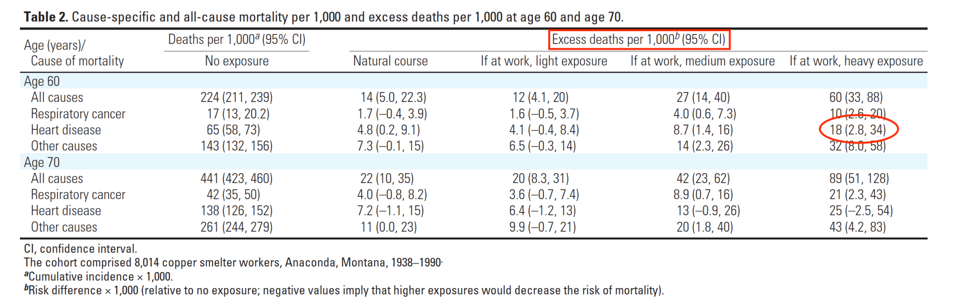


Table 2. Cause-specific and all-cause mortality per 1,000 and excess deaths per 1,000 at age 60 and age 70.

Age (years)/	Deaths per 1,000[a] (95% CI)	Excess deaths per 1,000[b] (95% CI)			
Cause of mortality	No exposure	Natural course	If at work, light exposure	If at work, medium exposure	If at work, heavy exposure
Age 60					
All causes	224 (211, 239)	14 (5.0, 22.3)	12 (4.1, 20)	27 (14, 40)	60 (33, 88)
Respiratory cancer	17 (13, 20.2)	1.7 (–0.4, 3.9)	1.6 (–0.5, 3.7)	4.0 (0.6, 7.3)	10 (2.6, 20)
Heart disease	65 (58, 73)	4.8 (0.2, 9.1)	4.1 (–0.4, 8.4)	8.7 (1.4, 16)	18 (2.8, 34)
Other causes	143 (132, 156)	7.3 (–0.1, 15)	6.5 (–0.3, 14)	14 (2.3, 26)	32 (8.0, 58)
Age 70					
All causes	441 (423, 460)	22 (10, 35)	20 (8.3, 31)	42 (23, 62)	89 (51, 128)
Respiratory cancer	42 (35, 50)	4.0 (–0.8, 8.2)	3.6 (–0.7, 7.4)	8.9 (0.7, 16)	21 (2.3, 43)
Heart disease	138 (126, 152)	7.2 (–1.1, 15)	6.4 (–1.2, 13)	13 (–0.9, 26)	25 (–2.5, 54)
Other causes	261 (244, 279)	11 (0.0, 23)	9.9 (–0.7, 21)	20 (1.8, 40)	43 (4.2, 83)

CI, confidence interval.

The cohort comprised 8,014 copper smelter workers, Anaconda, Montana, 1938–1990·

[a] Cumulative incidence × 1,000.

[b] Risk difference × 1,000 (relative to no exposure; negative values imply that higher exposures would decrease the risk of mortality).

Keil and Richardson (2017) *Env Health Persp*

Arsenic and cardiovascular disease in 8,014 copper smelter workers, reanalysis

To address healthy worker survivor bias, methods matter

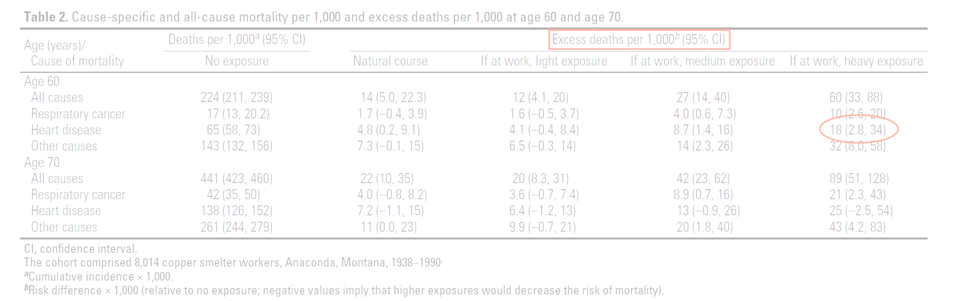


Table 2. Cause-specific and all-cause mortality per 1,000 and excess deaths per 1,000 at age 60 and age 70.

Age (years)/	Deaths per 1,000[a] (95% CI)	Excess deaths per 1,000[b] (95% CI)			
Cause of mortality	No exposure	Natural course	If at work, light exposure	If at work, medium exposure	If at work, heavy exposure
Age 60					
All causes	224 (211, 239)	14 (5.0, 22.3)	12 (4.1, 20)	27 (14, 40)	60 (33, 88)
Respiratory cancer	17 (13, 20.2)	1.7 (–0.4, 3.9)	1.6 (–0.5, 3.7)	4.0 (0.6, 7.3)	10 (2.6, 20)
Heart disease	65 (58, 73)	4.8 (0.2, 9.1)	4.1 (–0.4, 8.4)	8.7 (1.4, 16)	18 (2.8, 34)
Other causes	143 (132, 156)	7.3 (–0.1, 15)	6.5 (–0.3, 14)	14 (2.3, 26)	32 (8.0, 58)
Age 70					
All causes	441 (423, 460)	22 (10, 35)	20 (8.3, 31)	42 (23, 62)	89 (51, 128)
Respiratory cancer	42 (35, 50)	4.0 (–0.8, 8.2)	3.6 (–0.7, 7.4)	8.9 (0.7, 16)	21 (2.3, 43)
Heart disease	138 (126, 152)	7.2 (–1.1, 15)	6.4 (–1.2, 13)	13 (–0.9, 26)	25 (–2.5, 54)
Other causes	261 (244, 279)	11 (0.0, 23)	9.9 (–0.7, 21)	20 (1.8, 40)	43 (4.2, 83)

CI, confidence interval.

The cohort comprised 8,014 copper smelter workers, Anaconda, Montana, 1938–1990·

[a] Cumulative incidence × 1,000.

[b] Risk difference × 1,000 (relative to no exposure; negative values imply that higher exposures would decrease the risk of mortality).

Keil and Richardson (2017) *Env Health Persp*

Terminology note

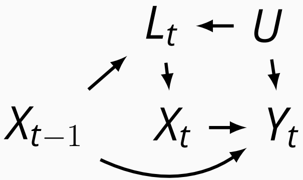
I use “**healthy worker survivor bias**” rather than “healthy worker survivor effect” because it is **a structural bias of specific parameters** that is amenable to control, rather than a singular effect.

To sharpen the distinction, it is helpful to see the structure

Concepts: the structure of healthy worker survivor bias

Healthy worker survivor bias: a structural problem

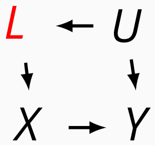
- DAG for HWSB



Buckley et al. (2015) *Epidemiology*

Healthy worker survivor bias: a structural problem

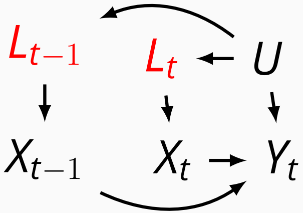
- Employment status = confounder (adjust)



Buckley et al. (2015) *Epidemiology*

Healthy worker survivor bias: a structural problem

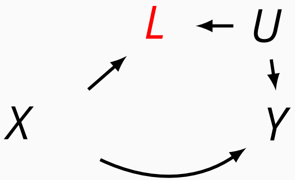
- Employment status = confounder (adjust)



Buckley et al. (2015) *Epidemiology*

Healthy worker survivor bias: a structural problem

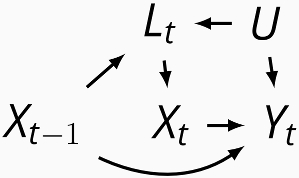
- Employment status = collider (don’t adjust)



Buckley et al. (2015) *Epidemiology*

Healthy worker survivor bias: a structural problem

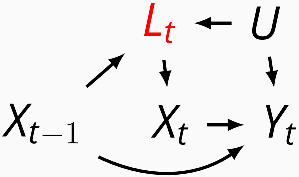
- Cumulative exposure = $\ldots + X_{t-1} + X_t$



Buckley et al. (2015) *Epidemiology*

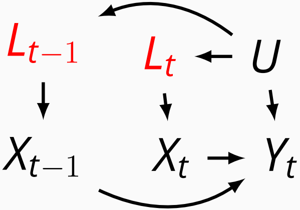
Healthy worker survivor bias: a structural problem

- Employment status = confounder (adjust)
- Employment status = collider (don't adjust)
- Cumulative exposure = $\ldots + X_{t-1} + X_t$
- Employment status = time-varying confounder affected by prior exposure

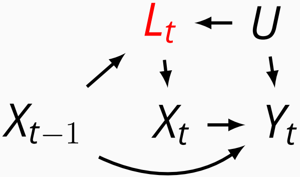


Buckley et al. (2015) *Epidemiology*

Healthy worker survivor bias: a structural problem



- **Healthy worker survivor bias** Confounding by employment status that *may* be complicated by impacts of exposure on employment



Buckley et al. (2015) *Epidemiology*

Healthy worker survivor bias: a structural problem summary

Confounding The bias to control

Collider stratification bias The bias to avoid when controlling confounding

Cumulative exposure We are interested in effects of multiple exposures at once, so we must concern ourselves with both of these biases at once

Methods for healthy worker survivor bias in cohort studies

Methods for healthy worker survivor bias in cohort studies

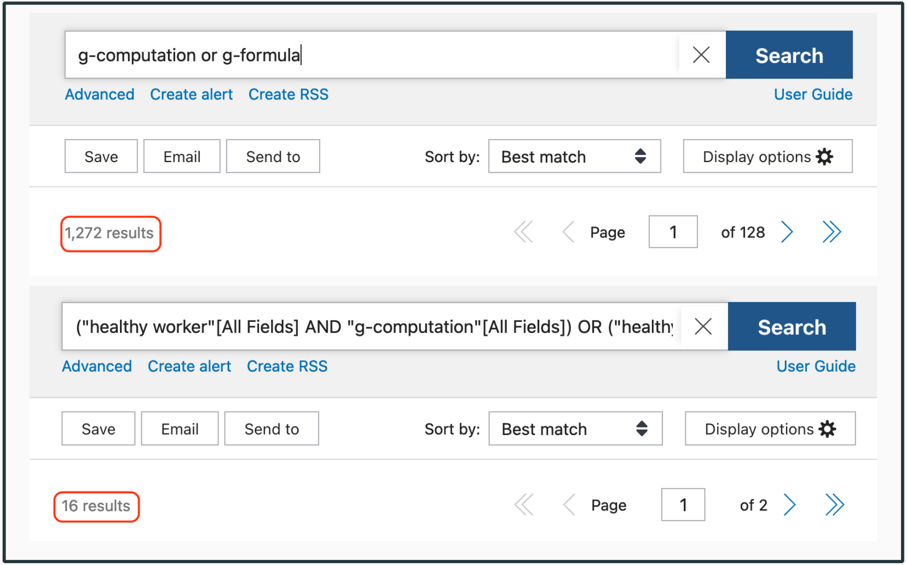
If exposure does not impact employment regression, g-methods[1]

If exposure impacts employment g-methods

G-methods g-estimation[2], g-formula/g-computation[3], inverse probability weighting[4]

[1] Naimi, Cole, and Kennedy (2017) *Int J Epidemiol*
[2] Keil, Richardson, and Troester (2015) *Am J Epidemiol*
[3] Keil et al. (2018) *Epidemiology*
[4] Keil et al. (2024) *Am J Epidemiol*

G-methods for healthy worker survivor bias in cohort studies



g-computation or g-formula

Sort by: Best match

1,272 results

Page 1 of 128

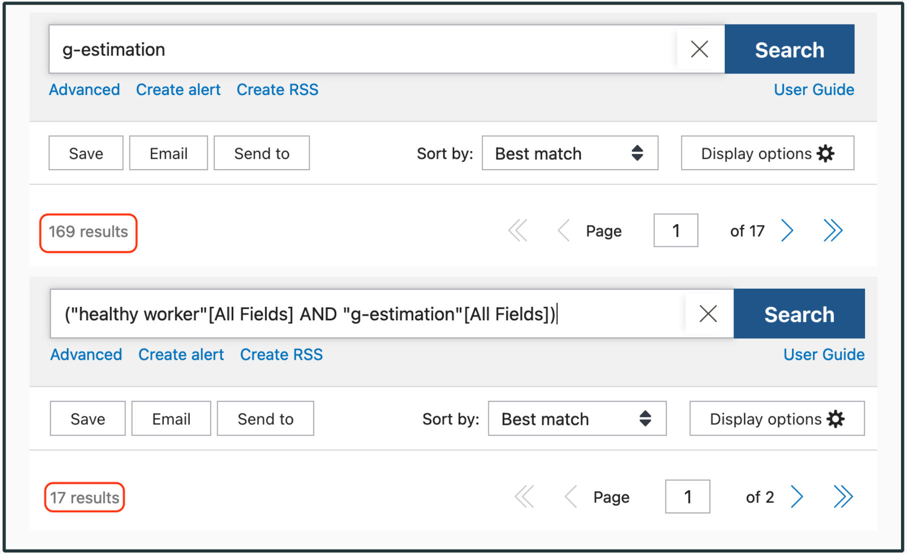
("healthy worker"[All Fields] AND "g-computation"[All Fields]) OR ("healthy

Sort by: Best match

16 results

Page 1 of 2

G-methods for healthy worker survivor bias in cohort studies



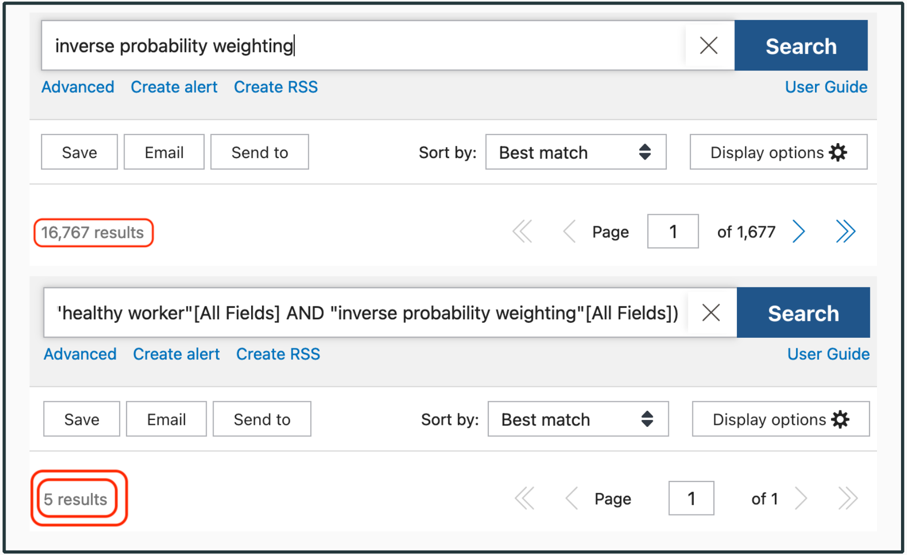
g-estimation

169 results

("healthy worker"[All Fields] AND "g-estimation"[All Fields])

17 results

G-methods for healthy worker survivor bias in cohort studies



inverse probability weighting

Sort by: Best match

16,767 results

Page 1 of 1,677

'healthy worker"[All Fields] AND "inverse probability weighting"[All Fields])

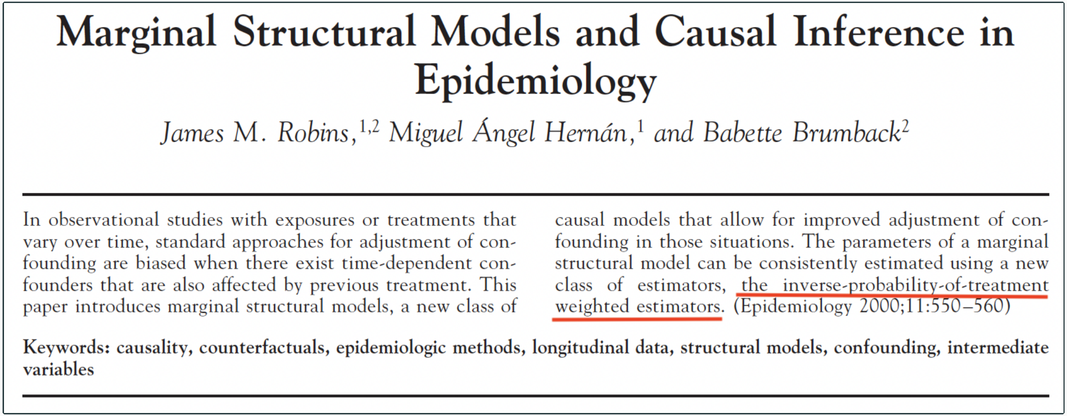
Sort by: Best match

5 results

Page 1 of 1

Compared with other g-methods, inverse probability weighting

- Is simpler
- Is less computationally demanding
- Reduces modeling assumptions
- Is conceptually clearer
- Is easier to interpret

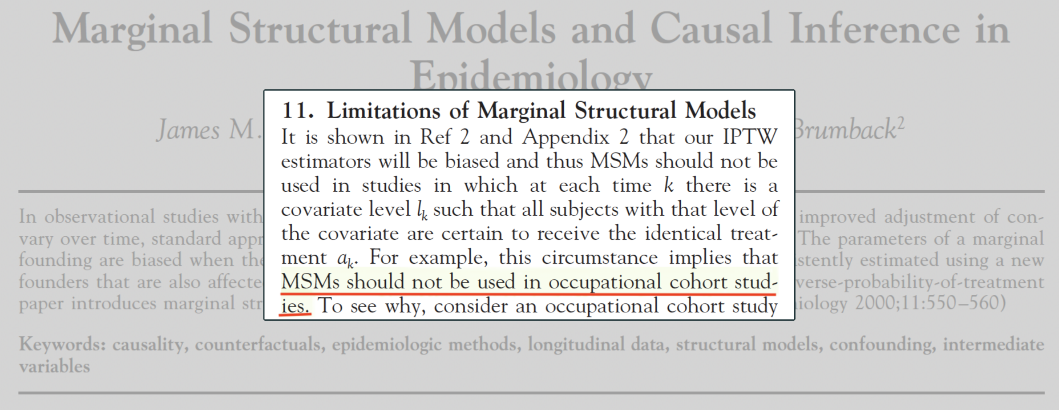


Marginal Structural Models and Causal Inference in Epidemiology

James M. Robins,[1,2] *Miguel Ángel Hernán,*[1] *and Babette Brumback*[2]

In observational studies with exposures or treatments that vary over time, standard approaches for adjustment of confounding are biased when there exist time-dependent confounders that are also affected by previous treatment. This paper introduces marginal structural models, a new class of causal models that allow for improved adjustment of confounding in those situations. The parameters of a marginal structural model can be consistently estimated using a new class of estimators, the inverse-probability-of-treatment weighted estimators. (Epidemiology 2000;11:550–560)

Keywords: causality, counterfactuals, epidemiologic methods, longitudinal data, structural models, confounding, intermediate variables



Marginal Structural Models and Causal Inference in Epidemiology

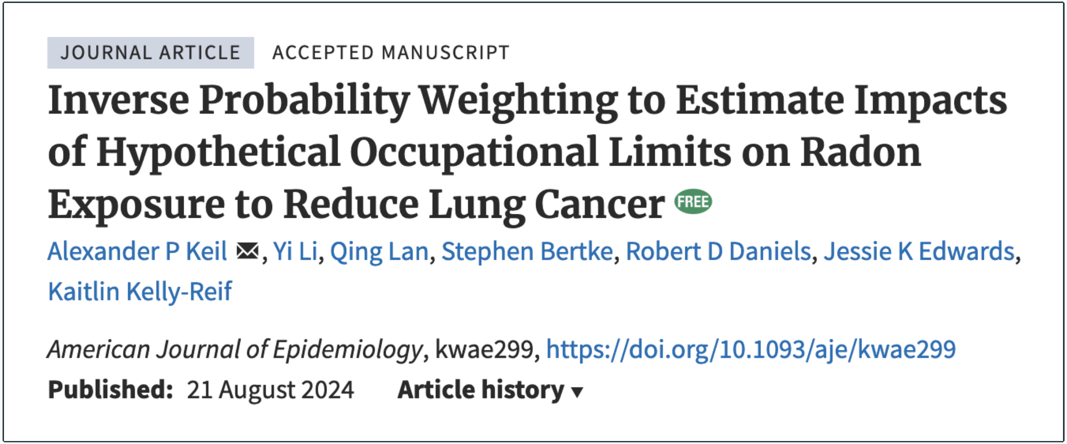
James M. … Brumback[2]

In observational studies with … improved adjustment of con-
vary over time, standard appr… The parameters of a marginal
founding are biased when the … stently estimated using a new
founders that are also affecte… verse-probability-of-treatment
paper introduces marginal str… iology 2000;11:550–560)

Keywords: causality, counterfactuals, epidemiologic methods, longitudinal data, structural models, confounding, intermediate variables

11. Limitations of Marginal Structural Models

It is shown in Ref 2 and Appendix 2 that our IPTW estimators will be biased and thus MSMs should not be used in studies in which at each time k there is a covariate level l_k such that all subjects with that level of the covariate are certain to receive the identical treatment a_k. For example, this circumstance implies that MSMs should not be used in occupational cohort studies. To see why, consider an occupational cohort study



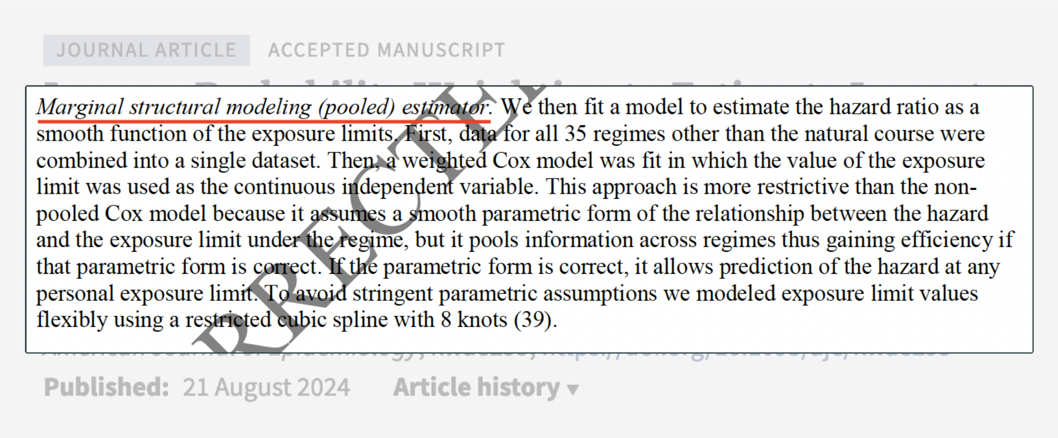
JOURNAL ARTICLE ACCEPTED MANUSCRIPT

Inverse Probability Weighting to Estimate Impacts of Hypothetical Occupational Limits on Radon Exposure to Reduce Lung Cancer FREE

Alexander P Keil ✉, Yi Li, Qing Lan, Stephen Bertke, Robert D Daniels, Jessie K Edwards, Kaitlin Kelly-Reif

American Journal of Epidemiology, kwae299, https://doi.org/10.1093/aje/kwae299

Published: 21 August 2024 **Article history** ▾



JOURNAL ARTICLE ACCEPTED MANUSCRIPT

Marginal structural modeling (pooled) estimator. We then fit a model to estimate the hazard ratio as a smooth function of the exposure limits. First, data for all 35 regimes other than the natural course were combined into a single dataset. Then, a weighted Cox model was fit in which the value of the exposure limit was used as the continuous independent variable. This approach is more restrictive than the non-pooled Cox model because it assumes a smooth parametric form of the relationship between the hazard and the exposure limit under the regime, but it pools information across regimes thus gaining efficiency if that parametric form is correct. If the parametric form is correct, it allows prediction of the hazard at any personal exposure limit. To avoid stringent parametric assumptions we modeled exposure limit values flexibly using a restricted cubic spline with 8 knots (39).

Published: 21 August 2024 **Article history ▾**

Technical note

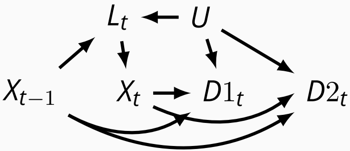
Hernán et al. and Cain et al. introduced the basic foundation for inverse probability weighting for *dynamic regimes*, which is a formal term for what I describe today.

Hernán et al. (2006) *Basic clin pharm & tox*
Cain et al. (2010) *The international journal of biostatistics*

A worked example with inverse probability weighting

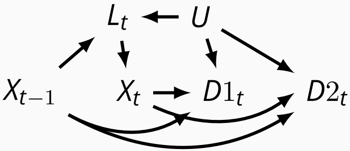
Synthetic example: effects of exposure in 10,000 worker cohort

- Full code available at https://github.com/alexpkeil1/2024_EPICOH
- Two outcomes : mimicking lung cancer ($D1$) and other-cause ($D2$) mortality based on US mortality rates (race/gender/age/year specific)
- One occupational exposure of interest (X) for chronic effects on lung cancer outcome $D1$
- Covariates: race, gender, age, year, wage status (salary vs. wage at hire), employment status L_t



Synthetic example: study questions

- What is the association (i.e. hazard ratio) between cumulative exposure and mortality, adusted for baseline exposure? (Cox model)
- What is the impact (i.e. hazard ratio) of a personal exposure limit on mortality (clone-censor-weight)
- What is the effect of any feasible personal exposure limit on mortality (clone-censor-weight + marginal structural model)



First 15 observations

```
id atwork     x  cumx wagestatus gender race age year d1 d2
 1      1 3.252 3.252          1      F    N  31 1956  0  0
 1      1 0.936 4.188          1      F    N  32 1957  0  0
 1      1 1.269 5.457          1      F    N  33 1958  0  0
 1      1 2.230 7.687          1      F    N  34 1959  0  0
 1      0 0.000 7.687          1      F    N  35 1960  0  0
 1      0 0.000 7.687          1      F    N  36 1961  0  0
 1      0 0.000 7.687          1      F    N  37 1962  0  0
 1      0 0.000 7.687          1      F    N  38 1963  0  0
 1      0 0.000 7.687          1      F    N  39 1964  0  0
 1      0 0.000 7.687          1      F    N  40 1965  0  0
 1      0 0.000 7.687          1      F    N  41 1966  0  0
 1      0 0.000 7.687          1      F    N  42 1967  0  0
 1      0 0.000 7.687          1      F    N  43 1968  0  0
 1      0 0.000 7.687          1      F    N  44 1969  0  0
 1      0 0.000 7.687          1      F    N  45 1970  0  0
```

Create variables for analyses (R code with `dplyr`)

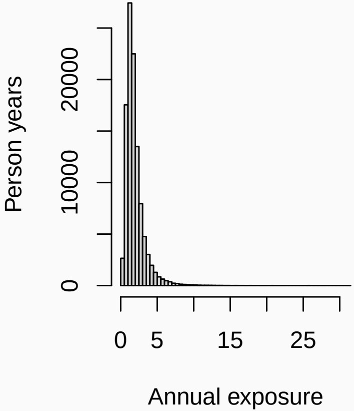
```
sim_cohort <- sim_cohort %>%
  group_by(id) %>%
  mutate(
    one = 1,
    time = cumsum(one),
    timein = time-1,
    agein = age-1,
    xl = lag(x, default=0),
    cxl = lag(cumx, default=0),
    cumatwork = cumsum(atwork),
    cumatworkl = lag(cumatwork, default=0),
    atworkl = lag(atwork, default=0),
    leftwork = as.numeric(atworkl==1 & atwork == 0),
    cxl2 = lag(cxl, default=0),
  ) %>%
  select(-one) %>%
  ungroup()
sim_cohort <- sim_cohort %>%
  group_by(id) %>%
  mutate(
  male = as.numeric(gender == 'M')
  ) %>%
  ungroup()
```

←Define helpful variables for time-dependent analysis (time on study, lagged variables)

Cohort characteristics at baseline, exposure and mortality by end of follow-up

N=10,000; 264,000 person-years

```
           Variable    N Pct Mean    SD
                Age               28  6.35
      Calendar year             1954 2.915
              Waged 7039  70
           Salaried 2961  30
           Gender F 5052  51
           Gender M 4948  49
         White race 7476  75
     Non-white race 2524  25
 Years of follow-up             26.4  16.9
Average X (at work)             1.89 0.797
       Cumulative X             17.2  15.3
       Years worked             9.07  7.67
                 D1  601   6
                 D2 7931  79
           Survived 1468  15
```



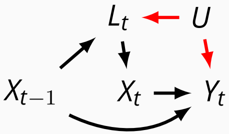
Should healthy worker survivor bias be a concern?

Should healthy worker survivor bias be a concern?

Is employment associated with mortality, given past exposures and confounders?

The hazard of $D1$ is lower for each additional year of employment

```
Call:
coxph(formula = Surv(agein, age, d1) ~ cumatwork + cxl + wagestatus +
    male + year, data = sim_cohort)

                coef exp(coef)  se(coef)      z       p
cumatwork  -0.008939  0.991100  0.014611 -0.612 0.54066
cxl         0.019431  1.019621  0.006887  2.821 0.00478
wagestatus  0.921821  2.513863  0.104730  8.802 < 2e-16
male        0.825716  2.283516  0.100563  8.211 < 2e-16
year       -0.001291  0.998710  0.006003 -0.215 0.82971

Likelihood ratio test=284.7  on 5 df, p=< 2.2e-16
n= 458152, number of events= 601
```

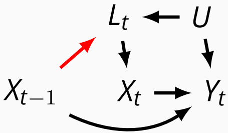


Should healthy worker survivor bias be a concern?

Is exposure associated with leaving work, given confounders?

Cumulative exposure is inversely related to leaving work (due to HWSB!)

```
Call:
coxph(formula = Surv(agein, age, leftwork) ~ cxl + wagestatus +
    male + year, data = filter(sim_cohort, atwork == 1 | leftwork ==
    1))

                coef  exp(coef)   se(coef)      z      p
cxl        0.0141371  1.0142375  0.0014321  9.872 <2e-16
wagestatus -0.0445361  0.9564411  0.0223938 -1.989 0.0467
male       -0.0057580  0.9942586  0.0220583 -0.261 0.7941
year       -0.0007632  0.9992371  0.0027040 -0.282 0.7778

Likelihood ratio test=220.8  on 4 df, p=< 2.2e-16
n= 115616, number of events= 9589
```

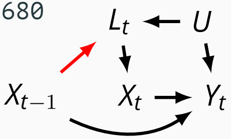


Should healthy worker survivor bias be a concern?

Is exposure associated with leaving work, given past exposures and confounders?

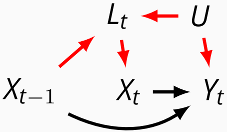
Recent exposure increases the hazard of leaving work

```
Call:
coxph(formula = Surv(agein, age, leftwork) ~ xl + cxl2 + wagestatus +
    male + year, data = filter(sim_cohort, atwork == 1 | leftwork ==
    1))

                coef exp(coef)  se(coef)      z        p
xl          0.071436  1.074049  0.006276 11.383  < 2e-16
cxl2        0.011064  1.011125  0.001480  7.476 7.65e-14
wagestatus -0.048541  0.952619  0.022392 -2.168   0.0302
male       -0.020670  0.979542  0.022081 -0.936   0.3492
year        0.002456  1.002459  0.002728  0.900   0.3680

Likelihood ratio test=298.4  on 5 df, p=< 2.2e-16
n= 115616, number of events= 9589
```



Should healthy worker survivor bias be a concern? Yes

We have evidence that the **components**[5] of healthy worker survivor bias are in place (time-varying confounder affected by prior exposure)



[5] Naimi et al. (2013) *Ann Epidemiol*

Clone-censor-weight approach to HWSB

Intuition of weighting

- All estimates can be conceptualized as a result of an experiment
- Observational data are not experiments
- By weighting observational data, we create a “pseudo-population”
- Each pseudo-population represents an arm of an experiment[6]

[6] Under causal assumptions, e.g. Hernán and Robins (2006) *Journal of Epidemiology & Community Health*

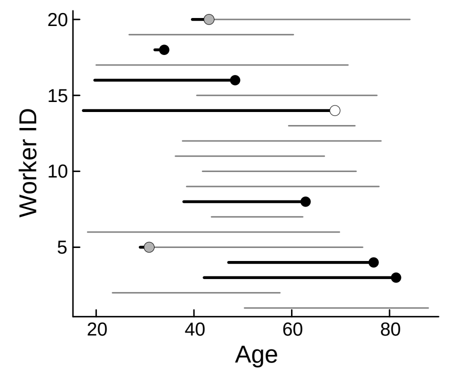
Clone-censor-weight approach to HWSB

Intuition of weighting

- "standard" inverse probability weighting for studying cumulative exposure-response curves *is not generally possible* for HWSB[7]
- Instead, we can use inverse probability weighting to study effects of hypothetical exposure limits[8]
- Example: "If at work, annual exposure can be no more than 2.0 units"[9]
- Sometimes called a "clone-censor-weight" approach

[7] Robins, Hernan, and Brumback (2000) *Epidemiology*
[8] Hernán et al. (2006) *Basic clin pharm & tox*
[9] Joffe (2012) *Epidemiology*

Clone-censor-weight approach to HWSB



From Keil et al. (2024) *Am J Epidemiol*: Person-time of 20 workers. **black line**: observed; **gray line**: observed but artificially censored; **gray dot**: artificially censored during follow-up; **black dot**: death while following regime; **white dot**: alive at end of study

Strategy:

Clone the cohort data for each hypothetical **exposure limit**

Censor Censor clones in when they exceed the limit

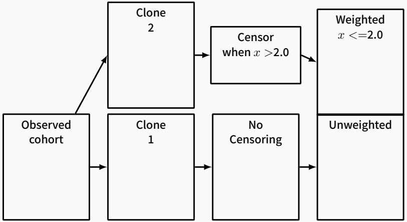
Weight Estimate weights based on the probability of *not* being censored

Weighted data $\approx$ cohort, had the exposure limit been in place

Clone-censor-weight approach to HWSB: Analysis

Estimating an effect of exposure

- The original data (with weights = 1) and the uncensored "clones" (with inverse-probability of censoring weights) are combined
- Set "exposed" to 1 in the clones and "exposed" to 0 in the original data
- Fit a weighted model (e.g. weighted Cox model) to the combined data with the "exposed" variable as the sole predictor



Clone-censor-weight approach to HWSB: Cloning

Make a copy of the cohort data and create a unique ID so the clones do not get confused with observed data

```
clones <- sim_cohort %>%
  mutate(
    limit = as.numeric(1.0),
    cloneid = paste0("clone2.0_", id)
  )
```

Clone-censor-weight approach to HWSB: Censoring

An individual is censored in the observation when a worker first exceeds an annual limit. Keep the observation in which censoring occurs (for now) and set “cens=1”.

```
cens_data <- clones %>%
  group_by(cloneid) %>%
  mutate(
    cens = pmin(1,cumsum(x > limit)),
    drop = pmin(2, cumsum(cens))
  ) %>%
  ungroup() %>%
  filter(
    drop < 2
  )
```

Clone-censor-weight approach to HWSB: predicting censoring

Separately model

- the “cens” variable at baseline (a “confounding” weight)
- the “cens” variable after baseline (a “censoring” weight)

```
# "censored at baseline" model
confdmod <- glm(cens~
  year + rcspline.eval(year, knots=yearkn0) +
  age + rcspline.eval(age, knots=agekn0) +
  wagestatus + male + race,
  data = cens_data, weight=conf_weight, family=binomial())

# "censored during follow-up" model
censdmod <- glm(cens~ cxl + cumatworkl +
  age + rcspline.eval(age, knots=agekn) +
  wagestatus + male + race , data = cens_data, weight=fu_weight,
  family=binomial())
```

Note: the weights used here are 1/0 weights that are just used to subset the data to baseline and non-baseline observations

Clone-censor-weight approach to HWSB: calculating estimated weights

The weight “ipw” is the time-specific, inverse, cumulative probability of *not* being censored and is set to zero when censoring occurs

```
cens_data$dconf = cens_data$conf_weight*as.numeric(predict(confdmod, type="response"))
cens_data$dcens = cens_data$fu_weight*as.numeric(predict(censdmod, type="response"))
cens_data$nconf = cens_data$conf_weight*as.numeric(predict(confnmod, type="response"))
cens_data$ncens = cens_data$fu_weight*as.numeric(predict(censnmod, type="response"))
cens_data <- cens_data %>%
  mutate(
    wtcontr = case_when(
      ((fobs____ == 1) & (atwork==1)) ~ (1-cens)*(1-nconf)/(1-dconf),
      ((fobs____ == 0) & (atwork==1)) ~ (1-cens)*(1-ncens)/(1-dcens),
      .default=1
    )
  ) %>%
  group_by(cloneid) %>%
  mutate(
    ipw = cumprod(wtcontr),
    ipwt = pmin(10, cumprod(wtcontr)) # sometimes truncated weights are advocated
  )
```

Clone-censor-weight approach to HWSB: Effect estimation

This code fits a cox model comparing the hazard under the “natural course” vs. the hazard under the intervention “limit annual exposures to 2.0 units at work”

```
sim_cohort$exposed = 0
cens_data$exposed = 1
sim_cohort$ipw = 1
sim_cohort$cloneid = paste0("cloneobs_", sim_cohort$id)
combined_data <- bind_rows(sim_cohort, cens_data)

coxph(Surv(agein, age, d1)~exposed,
     data=filter(combined_data, ipw != 0), weight=ipw,
     id=cloneid, cluster=id) %>%
      summary %>% print
```

Note: the “cluster” argument is necessary to get the robust variance, which accounts for the fact that individuals appear in the data more than once. Bootstrapping is another option.

Clone-censor-weight: more than 2 “interventions”

- Clone-censor-weight for exposure-response
- Combine many (30) weighted datasets with assigned exposure “level”
- Cox model with exposure level as only predictor
- “marginal structural model”

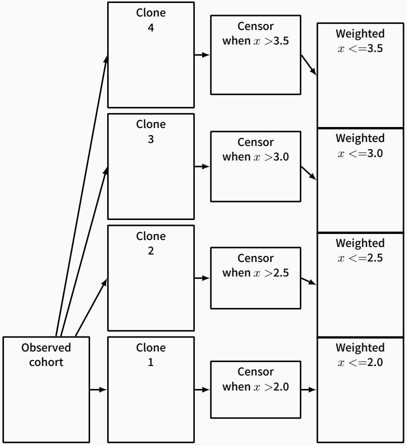


Figure: Conceptual model of clone-censor-weight approach with 4 interventions

Marginal structural models

This code fits a Cox model estimating the hazard ratio for a 1 unit increase in the occupational limit

```
ttr = coxph(Surv(agein, age, d1)~limit, data=filter(combined_wtd_data, ipw>0),
            id=cloneid, weight=ipw, cluster=id)
```

Model	Outcome	HR	(95% CI)	Interpretation
Adjusted Cox model, 1 unit increase in cumulative exposure	D1	1.015	(1.012, 1.019)	HR$>$1, exposure harmful
Clone/censor/weight, 2.0 limit vs. none (ref)	D1	0.12	(0.038, 0.37)	HR$<$1, exposure harmful
Clone/censor/weight, 1 unit increase in limit[10]	D1	1.098	(1.043, 1.16)	HR$>$1, exposure harmful

[10] Limits range from 2.2 to 5.7, with mean cumulative exposures ranging from 4 to 14

Discussion of HWSB

Summary

- Key: understanding structural components
- Bias often downward, if not through the null
- Cohort data are required for addressing
- G-methods like IPCW can address in general
- Example demonstrates a ”policy-response” model
- HWSB adjusted dose-response models are necessarily different

Discussion of inverse probability weighting

Limitations

- Can have high variance
- Sparse data = problematic
- Unfamiliar relative to mortality models with cumulative exposure

Strengths

- Addresses a key bias (HWSB)
- Simpler than other g-methods
- Weaker modeling assumptions
- Many published examples in other areas, e.g. with diagnostics
- Focuses effect estimates on worker health

References

Katie M Applebaum, Elizabeth J Malloy, and Ellen A Eisen. “Left truncation, susceptibility, and bias in occupational cohort studies”. In: *Epidemiology* 22.4 (2011), pp. 599–606.

Ove Björ et al. *Reduced mortality rates in a cohort of long-term underground iron-ore miners*. 2013.

Jessie P Buckley et al. “Evolving methods for inference in the presence of healthy worker survivor bias”. In: *Epidemiology* 26.2 (2015), pp. 204–212.

Lauren E Cain et al. “When to start treatment? A systematic approach to the comparison of dynamic regimes using observational data”. In: *The international journal of biostatistics* 6.2 (2010).

Miguel A Hernán and James M Robins. “Estimating causal effects from epidemiological data”. In: *Journal of Epidemiology & Community Health* 60.7 (2006), pp. 578–586.

Miguel A Hernán et al. “Comparison of dynamic treatment regimes via inverse probability weighting”. In: *Basic clin pharm & tox* 98.3 (2006), pp. 237–242.

I Hertz-Picciotto, SW Hu, and HM Arrighi. “Re:” Does arsenic exposure increase the risk for circulatory disease?”-The authors’ reply”. In: *Am J Epidemiol* 152.3 (2000), pp. 293–293.

Marshall M Joffe. “Commentary: Structural Nested Models, G-Estimation, and the Healthy Worker Effect: The Promise (Mostly Unrealized) and the Pitfalls”. In: *Epidemiology* 23.2 (2012), pp. 220–222.

Alexander P Keil and David B Richardson. “Reassessing the link between airborne arsenic exposure among anaconda copper smelter workers and multiple causes of death using the parametric g-formula”. In: *Env Health Persp* 125.4 (2017), pp. 608–614.

Alexander P Keil, David B Richardson, and Melissa A Troester. “Healthy worker survivor bias in the Colorado Plateau uranium miners cohort”. In: *Am J Epidemiol* 181.10 (2015), pp. 762–770.

Alexander P Keil et al. “Estimating the impact of changes to occupational standards for silica exposure on lung cancer mortality”. In: *Epidemiology* 29.5 (2018), pp. 658–665.

Alexander P Keil et al. “Inverse probability weighting to estimate impacts of hypothetical occupational limits on radon exposure to reduce lung cancer”. In: *Am J Epidemiol* (2024), kwae299.

Jay H Lubin and Joseph F Fraumeni Jr. “Re:“Does arsenic exposure increase the risk for circulatory disease?”” In: *Am J Epidemiol* 152.3 (2000), pp. 290–293.

Ashley I Naimi, Stephen R Cole, and Edward H Kennedy. “An introduction to g methods”. In: *Int J Epidemiol* 46.2 (2017), pp. 756–762.

Ashley I Naimi et al. “Assessing the component associations of the healthy worker survivor bias: occupational asbestos exposure and lung cancer mortality”. In: *Ann Epidemiol* 23.6 (2013), pp. 334–341.

Andreas M Neophytou et al. “Marginal structural models in occupational epidemiology: application in a study of ischemic heart disease incidence and PM2. 5 in the US aluminum industry”. In: *Am J Epidemiol* 180.6 (2014), pp. 608–615.

William Ogle. “Letter to the Registrar-General on the mortality in the registration districts of England and Wales during the ten years 1871–80”. In: *Supplement to the Forty-Fifth Annual Report of the Registrar-General of Births, Deaths, and Marriages in England* (1885).

Neil Pearce, Harvey Checkoway, and David Kriebel. “Bias in occupational epidemiology studies”. In: *Occ Env Med* 64.8 (2007), pp. 562–568.

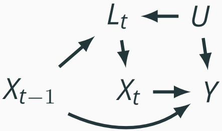
James M Robins, Miguel Angel Hernan, and Babette Brumback. *Marginal structural models and causal inference in epidemiology*. 2000.

Concepts and Methods for Healthy Worker Survivor Bias

Alex Keil

December 2024

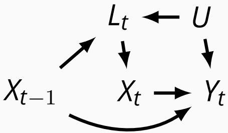
Earl Stadtman Investigator, National Cancer Institute



Extra topics

Employment status is a lever, but not the only one

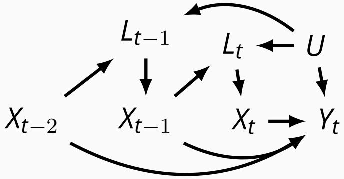
- Bias is ultimately confounding by underlying health status
- If health status can be measured directly, then employment status is not needed[11]
- Other factors to consider: job title, work area/tasks[12]



[11] e.g. Neophytou et al. (2014) *Am J Epidemiol*
[12] Pearce, Checkoway, and Kriebel (2007) *Occ Env Med*

Selection issues and selection bias

Examples

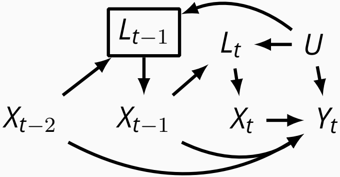


[13] Applebaum, Malloy, and Eisen (2011) *Epidemiology*
[14] e.g. Keil, Richardson, and Troester (2015) *Am J Epidemiol*

Selection issues and selection bias

Examples

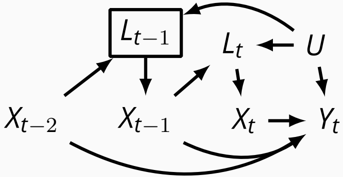
- **Bias from left truncation** Population selected among prevalent workers at a given time; incident hires[13] or use modified cumulative exposure[14]
- **Depletion of susceptibles** U will be an effect measure modifier and prevalent workers will be healthier/less susceptible (external validity)

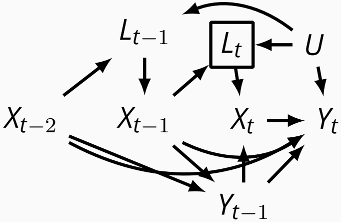


[13] Applebaum, Malloy, and Eisen (2011) *Epidemiology*
[14] e.g. Keil, Richardson, and Troester (2015) *Am J Epidemiol*

Selection issues and selection bias

Examples

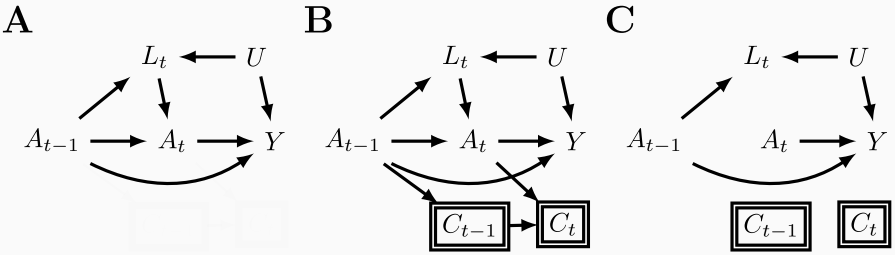




- **Reverse causation** Cross sectional studies of workers, restrict to only recent exposures

[13] Applebaum, Malloy, and Eisen (2011) *Epidemiology*
[14] e.g. Keil, Richardson, and Troester (2015) *Am J Epidemiol*

Clone-censor-weight for a single occupational annual limit



A Cloned (same as observed data)

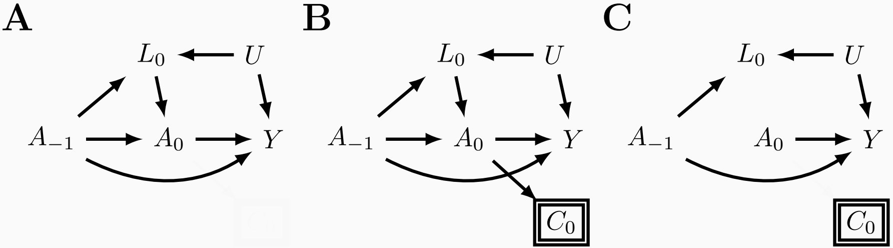
B Censored (remove observations from $t - T$ if exceed limit in time t)

C Weighted ("intervention" on censoring, a deterministic node)

Keil et al. (2024) *Am J Epidemiol*

Artificial “censoring” at baseline

Note: mortality follow-up began after some workers had been exposed



A Cloned (same as observed data)

B Censored (remove worker if exceed limit in first observation)

C Weighted (“intervention” on entry into pseudo-cohort)

Keil et al. (2024) *Am J Epidemiol*

Inverse probability weighting to estimate effects of occupational limits

- IPW is much less reliant on modeling assumptions
- Could be a routinely implemented (contrasted with other g-methods)
- More work: specifying a policy MSM for more principled extrapolation