HARVARD T.H. CHAN | SCHOOL OF PUBLIC HEALTH
Powerful ideas for a healthier world

Analysis of case-cohort study with the additive hazards model

Kate Hu

April, 2024

European Journal of Epidemiology (2021) 36:1129–1142
https://doi.org/10.1007/s10654-021-00739-3

METHODS

Estimating the hazard rate difference from case-cohort studies

Jie K. Hu[1] · Kwun C. G. Chan[1] · David J. Couper[2] · Norman E. Breslow[1]

Topics

- Additive Hazards Models
- Case-Cohort Studies
- Analysis R code[3]
  - Construct Weight
  - Handle Missing Covariates
  - Use Auxiliary Information to Improve Precision
- Biomarkers
- Open-source exercise code and datasets for practice

Links

- http://www.katehu.com/proxies/

Scientific Questions

- Whether elevated levels of high sensitivity C-reactive protein (hs-CRP) is associated with the increased risk of Coronary Heart Disease (CHD)

- Individual risk prediction based on traditional risk factors and the new biomarker hs-CRP together, particularly among people with low density lipoprotein cholesterol (LDL-C)

- Impact: hs-CRP may Identify some patients traditional risk factor measurements could not identify for subsequent preventive therapies

Background

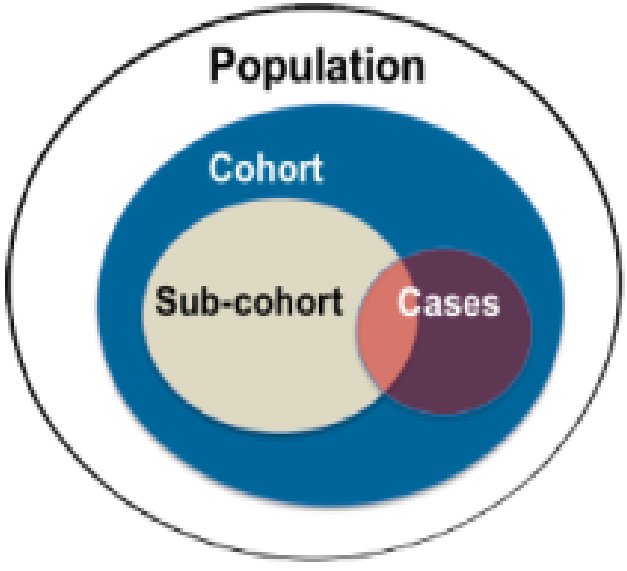
- Atherosclerosis Risk in Communities Study (ARIC) study (ARIC Investigators, 1989):
  - prospective epidemiologic study
  - to investigate the causes, the outcomes and the risk factors related to cardiovascular diseases

- A Biomarker sub-study (Ballantyne et al. 2004)
  - hs-CRP and Lp-PLA2 were assessed for a subset

Datasets

- The main cohort was followed for a CHD event and measured for traditional risk factors
- A cohort random sample (CRS) was selected using stratified sampling
  - sex
  - race
  - age
- hs-CRP were assessed for the CRS members and the subsequently identified CHD cases using stored plasma

Survival Analysis for Case–Cohort Studies



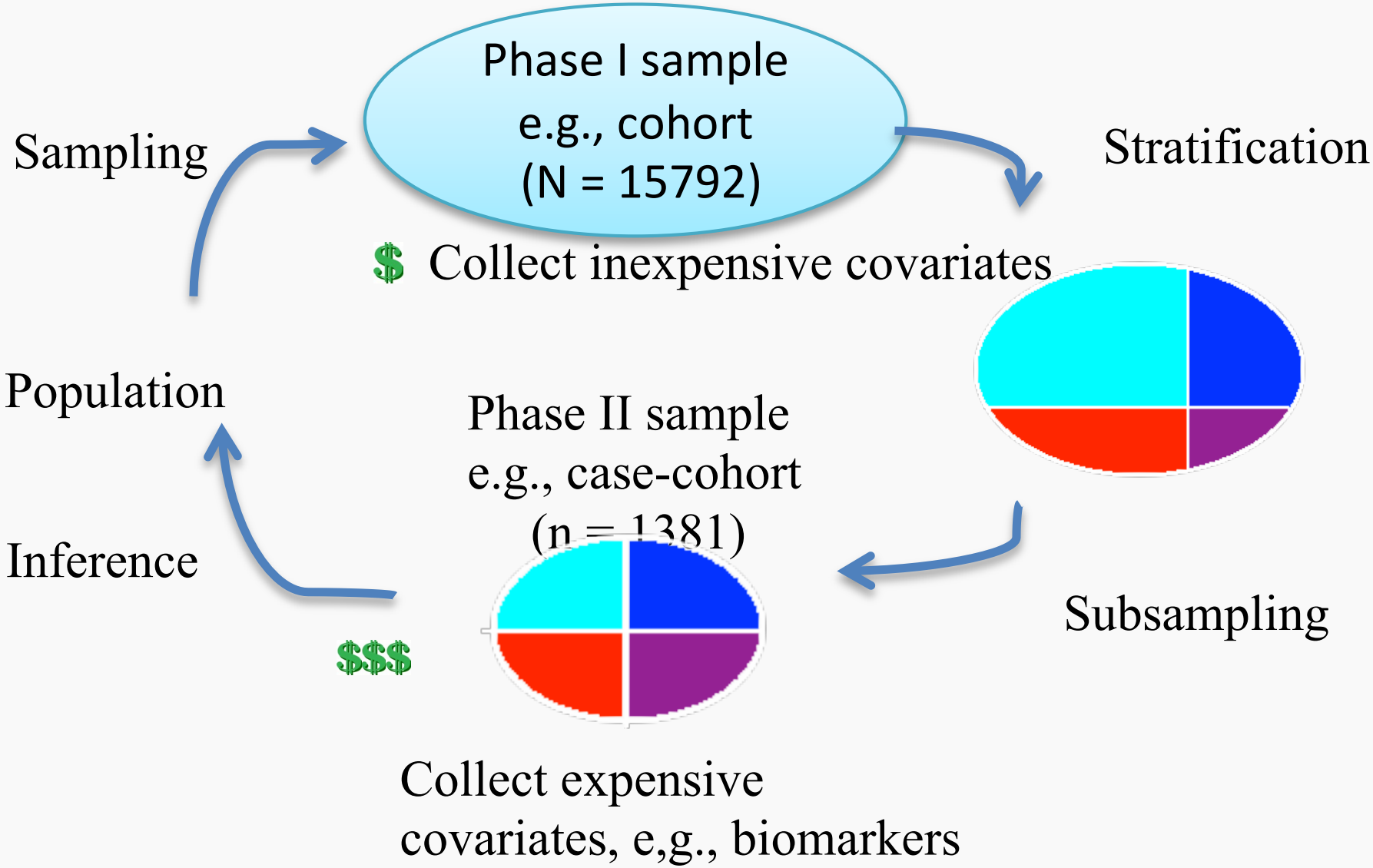
Aim
investigate the association between a biomarker and a disease

Study Design

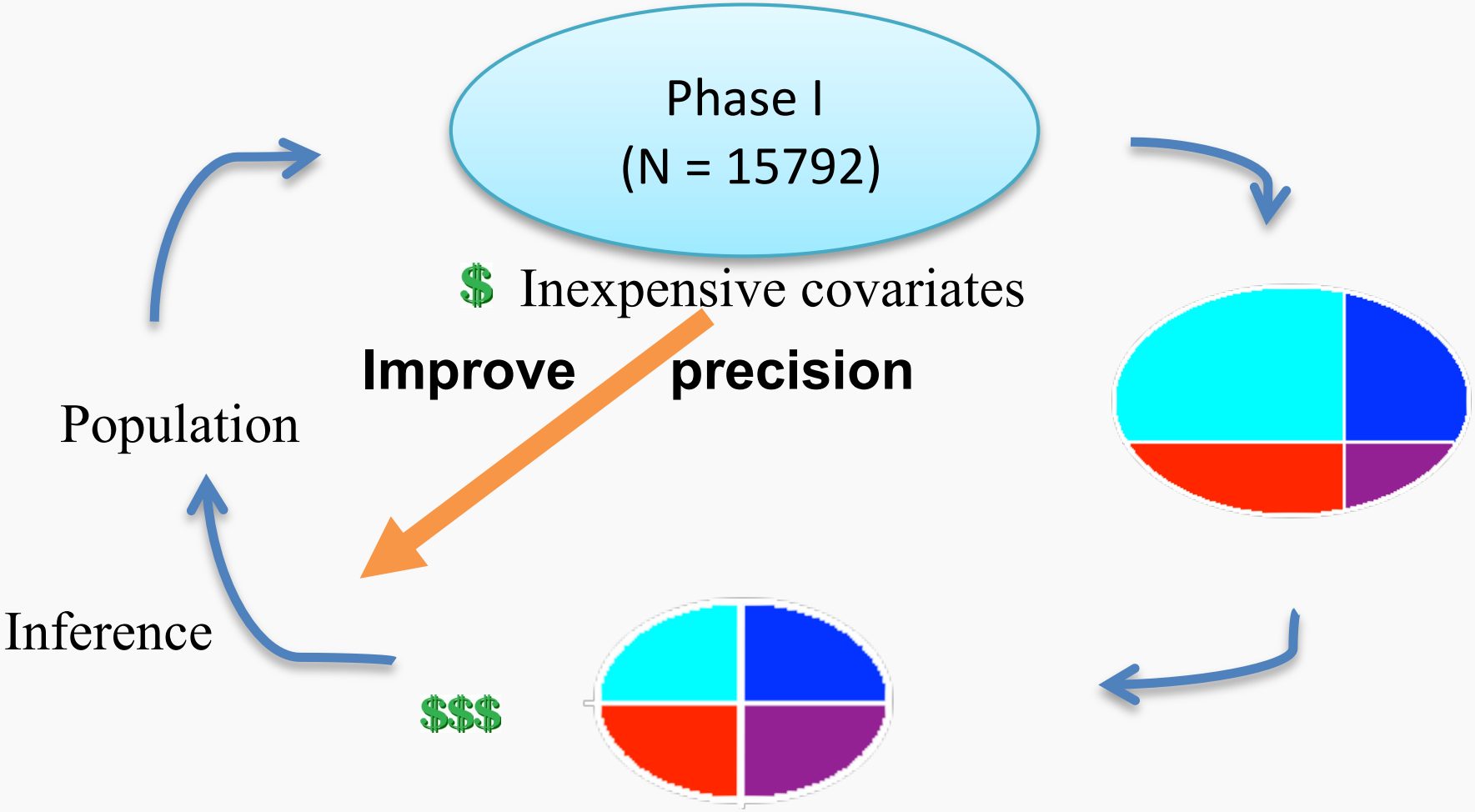
Auxiliary covariates collected for Cohort: 15792 participants

Biomarker measurements for { Subcohort: 777 members; Case: 604 incidents

Two-Phase Sampling Design Reduces Study Costs
A Sample Survey Approach to Analyzing Case-Cohort Studies

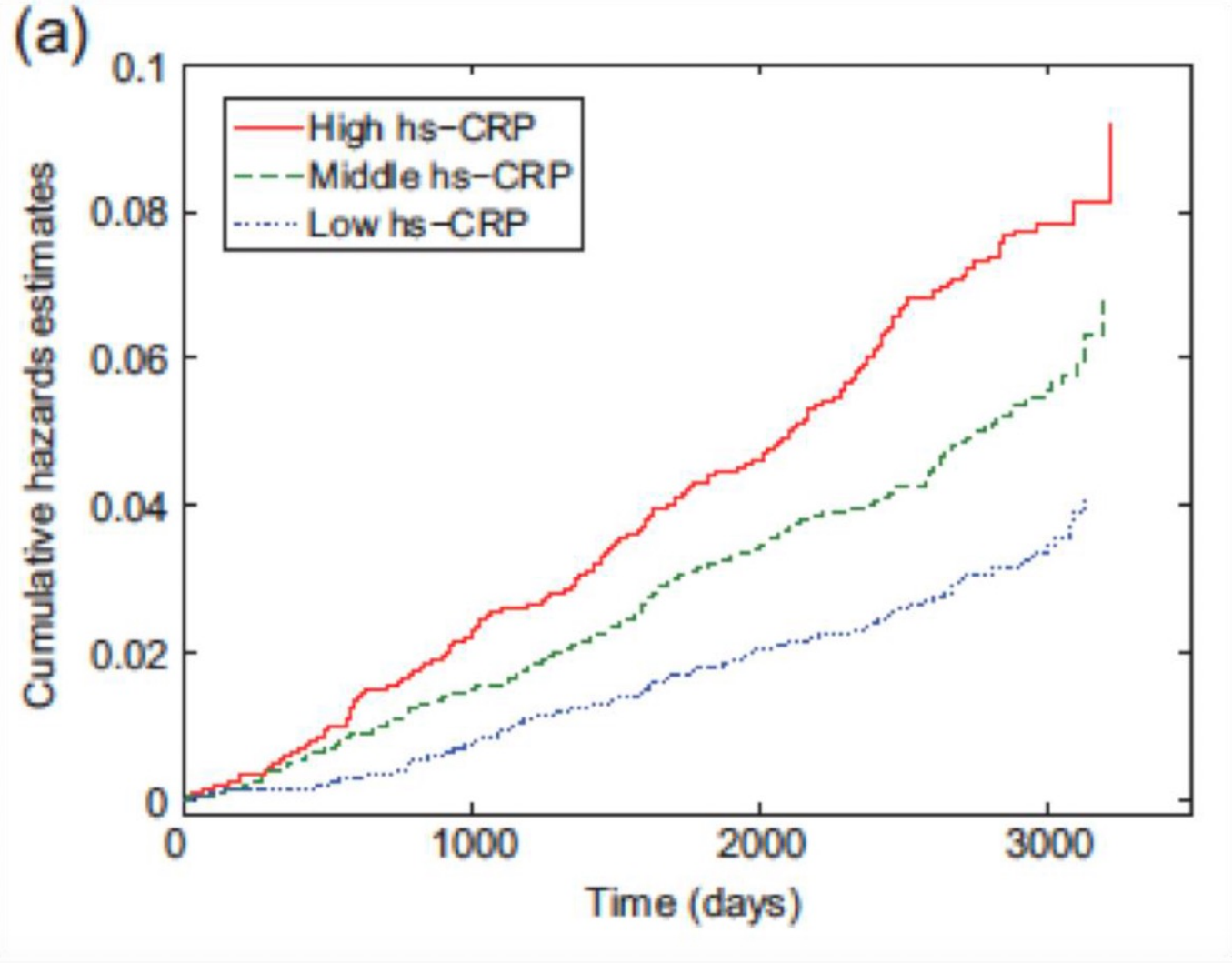


Incorporate Auxiliary Information to Improve Efficiency in Two-Phase Sampling Studies



Additive Hazards Model

$$\lambda(t|z) = \lambda_0(t) + z^T\theta$$



$$\Lambda(t|z=2) - \Lambda(t|z=1) = \theta t$$

(Lin & Ying, 1994)
(Kang, et.al, 2013)

Benefits of the Additive Hazards Models

- the absolute effect in the unit of cases per person-time of observation
- more relevant than the relative risk when evaluating the public health impact of an intervention or a risk factor
- unlike the hazard ratio (HR) the hazard difference (HD) is transportable across study populations and the conditional effect obtained by the AH model is equal to the marginal effect
- a constant HR assumed by the Cox model for each cause-specific hazard does not guarantee a constant HR for all-cause mortality while this inconsistency is not present for the HD.
- the most natural for measuring interaction effects

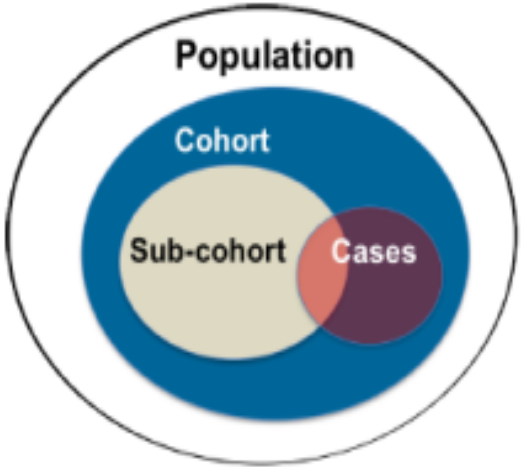
R Package addhazard for Fitting the Additive Hazards Model to Case–Cohort Study Data

Model 1: $$\lambda(t|z) = \lambda(t) + \beta_1 * \text{hs-CRP}_m + \beta_2 * \text{hs-CRP}_h + \beta_3 * \text{SEX} + \beta_4 * \text{AGE} + \beta_5 * \text{RACE}$$

where hs-CRP_m and hs-CRP_h are binary variables indicating whether an individual belongs to the medium and high-level groups of hs-CRP. The R commands to fit this model and retrieve its result are:

```
SEED = 20
model1 <- Surv(SURVTIME,CHD)~crp+AGE+SEX+RACE
fit1 <- ah.2ph(model1, R = in_phase2, weights = w, data = aric,
               robust = FALSE, ties = "break", seed = SEED)
summary(fit1)
```

Weights Construction




```
SEED = 20
model1 <- Surv(SURVTIME,CHD)~crp+AGE+SEX+RACE
fit1 <- ah.2ph(model1, R = in_phase2, weights = w, data = aric,
               robust = FALSE, ties = "break", seed = SEED)
summary(fit1)
```

```
aric$w[aric$CHD==1] <- sum(aric$INCRS==1 & aric$CHD==1)
            /  sum(aric$INCRS==1 & aric$CHD==1 & (!is.na(aric$crp)))
aric$w[aric$CHD==0] <- aric$WGT1[aric$CHD==0]
```

Results: Weights

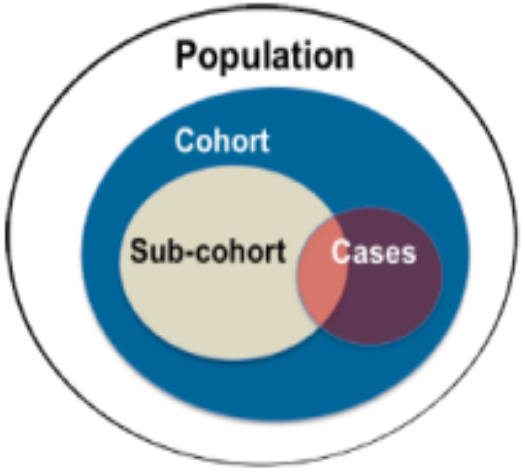


Table 1 Variable probability sampling weights for the ARIC biomarker substudy dataset

	CHD controls								CHD cases
	Black				White				
	Female		Male		Female		Male		
	Age ≥ 55	Age <55	Age ≥ 55	Age <55	Age ≥ 55	Age <55	Age ≥ 55	Age <55	
Stratum	1	2	3	4	5	6	7	8	9
Weights	14.3	19.5	6.1	16.1	15.1	32.0	12.3	17.7	1.21

R Code to Easily Incorporate Auxiliary Variables

```
SEED = 20
model3 <- Surv(SURVTIME, CHD)~BMI+TRIG+AGE+SEX+RACE
fit3 <- ah(model3, data = aric, ties = "break", seed = SEED)
aric$res_BMI <- fit3$resid[, "BMI"]
aric$res_TRIG <- fit3$resid[, "TRIG"]
fit1_calibrated<-ah.2ph(model1, R = in_phase2, weights = w, data = aric,
                        robust = FALSE, ties = "break", seed = SEED,
                        calibration.variables = c("res_BMI", "res_TRIG"))
summary(fit1_calibrated)
```

Calibration variables to consider: age, age^2, continuous outcome variable *T,* the binary outcome variables Δ, etc.

Results: Precision for Estimators Improves

$$\lambda(t|z) = \lambda_0(t) + z^T\theta$$

Table 2 CHD HDs per 1000 person-years (95% CI) by hs-CRP levels and their model-based standard errors with and without calibration

From: Estimating the hazard rate difference from case-cohort studies

	Standard weights			Calibrated weights		
	HD (95% CI)	SE	*p* value	HD (95% CI)	SE	*p* value
Model 1[a]						
hs-CRP 1.0–3.0 mg/L[c]	3.05 (1.16–4.95)	0.97	0.0016	3.02 (1.18–4.87)	0.94	0.0013
hs-CRP >3.0 mg/L	7.00 (4.61–9.39)	1.22	< 0.0001	7.20 (4.89–9.51)	1.18	< 0.0001
Model 2[b]						
hs-CRP 1.0–3.0 mg/L	1.40 (– 0.72–3.52)	1.08	0.1958	1.41 (– 0.69–3.51)	1.07	0.1879
hs-CRP > 3.0 mg/L	4.58 (2.06–7.10)	1.28	0.0004	4.75 (2.27–7.23)	1.26	0.0002
Model 2, LDL-C < 130 mg/dL						
hs-CRP 1.0–3.0 mg/L	– 0.01 (– 2.36–2.33)	1.20	0.9902	0.03 (– 2.37–2.43)	1.23	0.9809
hs-CRP > 3.0 mg/L	2.70 (– 0.14–5.54)	1.45	0.0624	2.94 (0.03–5.85)	1.49	0.0475

[a]Adjusted for age, sex, and race

[b]Adjusted for age, sex, race, smoking status, systolic blood pressure, LDL-C, HDL-C, and diabetes

[c]hs-CRP <1.0 mg/L is the reference group

Results: Auxiliary Variable Choice

Table 3 Standard errors of HDs using different calibration variables (cases per 1000 person-years)

	SE 0	SE I	SE II	SE III	SE IV
hs-CRP 1.0− 3.0 mg/L	1.080	1.072	1.124	1.095	1.147
hs-CRP > 3.0 mg/L	1.285	1.264	1.331	1.321	1.326
Age	0.093	0.093	0.085	0.095	0.064
Sex (male)	1.171	1.173	0.994	1.158	0.786
Race (white)	1.173	1.145	1.056	1.226	0.813
Smoking (former vs. current)	1.584	1.581	1.648	1.567	1.020
Smoking (never vs. current)	1.475	1.478	1.526	1.465	0.967
Systolic blood pressure (SBP)	0.033	0.032	0.034	0.035	0.023
LDL-C	0.016	0.016	0.017	0.017	0.011
HDL-C	0.028	0.027	0.029	0.029	0.018
Diabetes	1.939	1.881	2.022	2.192	1.312

SE 0: no calibration

SE I: calibration variables are the integrated martingale residuals for covariates BMI and triglycerides obtained from fitting model 4 to the phase I sample

SE II: calibration variables are strata indicators

SE III: calibration variables are the baseline variables including age, sex, race, smoking status, diabetes, SBP, LDL-C and HDL-C

SE IV: calibration variables are the integrated martingale residuals for all the above baseline variables obtained from fitting model 4 to the phase I sample

Proxies are Weak

Table 4 Weighted correlation coefficients ρ between hs-CRP and phase I cohort variables

	hs-CRP (continuous)	hs-CRP (1.0–3.0 mg/L)	hs-CRP (> 3.0 mg/L)
Body Mass Index (BMI)	0.431	0.293	0.461
Triglycerides	0.142	0.267	0.297
Hypertension history	0.208	0.174	0.274
Diabetes	0.203	0.074	0.230
Systolic blood pressure (SBP)	0.154	0.199	0.252
Diastolic blood pressure (DBP)	0.044	0.108	0.114
Total cholesterol	0.029	0.130	0.096
LDL-C	0.007	0.069	0.057
HDL-C	− 0.050	− 0.052	− 0.093
Smoking (never vs. current)	− 0.010	− 0.078	− 0.058
Smoking (former vs. current)	− 0.040	0.030	− 0.064

Results: Auxiliary Variable Choice

Table 3 Standard errors of HDs using different calibration variables (cases per 1000 person-years)

	SE 0	SE I	SE II	SE III	SE IV
hs-CRP 1.0− 3.0 mg/L	1.080	1.072	1.124	1.095	1.147
hs-CRP > 3.0 mg/L	1.285	1.264	1.331	1.321	1.326
Age	0.093	0.093	0.085	0.095	0.064
Sex (male)	1.171	1.173	0.994	1.158	0.786
Race (white)	1.173	1.145	1.056	1.226	0.813
Smoking (former vs. current)	1.584	1.581	1.648	1.567	1.020
Smoking (never vs. current)	1.475	1.478	1.526	1.465	0.967
Systolic blood pressure (SBP)	0.033	0.032	0.034	0.035	0.023
LDL-C	0.016	0.016	0.017	0.017	0.011
HDL-C	0.028	0.027	0.029	0.029	0.018
Diabetes	1.939	1.881	2.022	2.192	1.312

SE 0: no calibration

SE I: calibration variables are the integrated martingale residuals for covariates BMI and triglycerides obtained from fitting model 4 to the phase I sample

SE II: calibration variables are strata indicators

SE III: calibration variables are the baseline variables including age, sex, race, smoking status, diabetes, SBP, LDL-C and HDL-C

SE IV: calibration variables are the integrated martingale residuals for all the above baseline variables obtained from fitting model 4 to the phase I sample

Summary

improvement in estimation precision by calibration is very specific—we see improvement only for the explanatory variables that are related to the calibration variables.

not only the strength of the relationship between the calibration variables and the explanatory variable matters for improving estimation precision but also how these variables should be used

Insights: Information About the Covariance Matters

$$\epsilon \equiv \int_0^\tau \left\{ \mathbf{Z}' - \frac{EY(t)\mathbf{Z}'}{EY(t)} \right\} \left\{ dN(t) - Y(t)d\hat{\Lambda}^*(t) - \mathbf{Z}'^T \hat{\theta}^* dt \right\}$$

$$(\mathbf{Z}' - E(\mathbf{Z}'))(Y - E(Y))$$

Insights: Theoretical Results on Asymptotic Variance

$$Var_A\left[\dot{\psi}_0\sqrt{N}(\hat{\alpha}-\alpha_0)h\right] = P\psi^2_{\alpha_0,h}$$

$$Var_A\left[\dot{\psi}_0\sqrt{N}(\hat{\alpha}^*-\alpha_0)h\right] = P\psi^2_{\alpha_0,h} + Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}\psi^2_{\alpha_0,h}\right]$$

$$Var_A\left[\sqrt{N}\dot{\psi}_0(\hat{\alpha}^{**}-\alpha_0)h\right] = P\psi^2_{\alpha_0,h} + Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}\{\psi_{\alpha_0,h}-\Pi(\psi_{\alpha_0,h}|\tilde{V})\}^2\right]$$

$\Pi(\cdot|\tilde{V})$ refers to population least squares projection on the space spanned by the calibration variables $\tilde{V}$:

$$\Pi(\cdot|\tilde{V}) = Q\{\cdot\tilde{V}^T\}(Q\tilde{V}\tilde{V}^T)^{-1}\tilde{V}$$

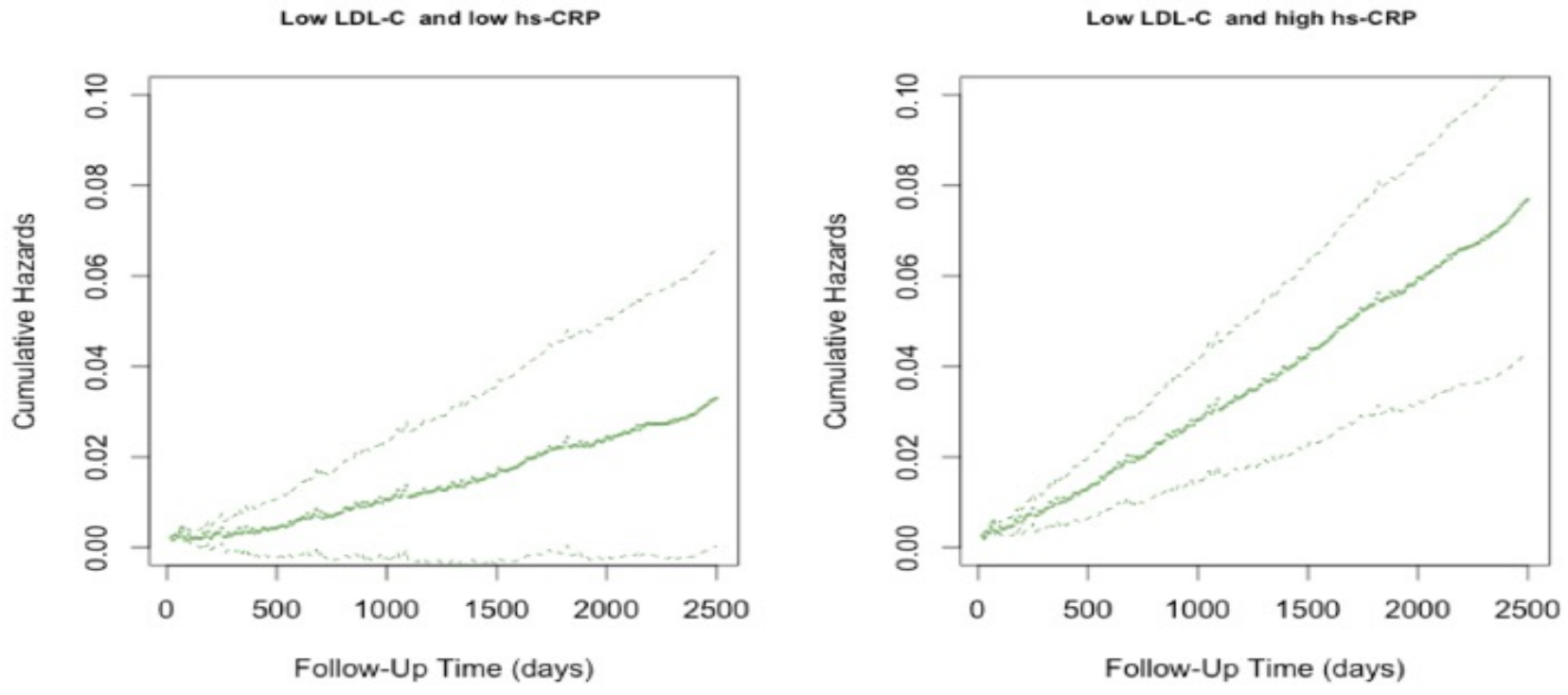
Related Topics

- Estimation Methods
- Calibration Techniques
- Risk Prediction[24]
- Simulation Results
- Finite population stratified sampling
- Interaction effects of Lp-PLA2 and hs-CRP

Can Use the Biomarker to Identify and Visualize High-risk Individuals

$$\lambda(t|z) = \lambda(t) + z^T\theta$$

The risk profile of patients with different hs-CRP and low density lipoprotein-cholesterol (LDL-C) levels



Jie Hu. *University of Washington ResearchWorks Archive* PhD Diss.

Public Datasets to Practice: National Wilms Tumor Study

R library addhazard

?nwts2ph

Examples for 'addhazard::ah.2ph'

Fit Additive Hazards Regression Models to Two-phase Sampling

Aliases: ah.2ph

Keywords:

```
### ** Examples

library(survival)
### fit an additive hazards model to two-phase sampling data without calibration
nwts2ph$trel <- nwts2ph$trel + runif(dim(nwts2ph)[1],0,1)*1e-8
fit1 <- ah.2ph(Surv(trel,relaps) ~ age + histol, ties = FALSE, data = nwts2ph, R = in.ph2, Pi = Pi,
 robust = FALSE,  calibration.variables = NULL)
summary(fit1)
```

```
Call:
ah.2ph(formula = Surv(trel, relaps) ~ age + histol, data = nwts2ph,
    R = in.ph2, Pi = Pi, ties = FALSE, robust = FALSE, calibration.variables = NULL)

          coef        se  lower.95  upper.95     z  p.value
[1,] 0.0020173 0.0004377 0.0011594 0.0028751 4.609 4.05e-06 ***
[2,] 0.0455141 0.0052445 0.0352349 0.0557933 8.678  < 2e-16 ***
```

Public Datasets to Practice: Breast Cancer dataset

Data:
https://www.mn.uio.no/math/english/research/groups/statistics-data-science/handbook-of-case-control-studies/chapter-17/

Tutorial:
https://www.mn.uio.no/math/english/research/groups/statistics-data-science/handbook-of-case-control-studies/chapter-17/bc_ah_analysis_for_table_17.4.html

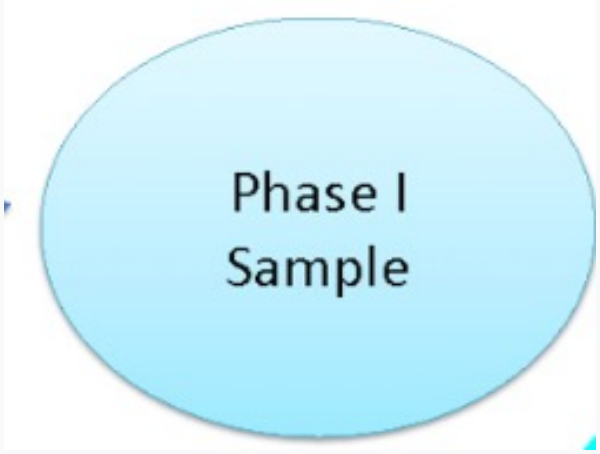
http://www.katehu.com/proxies/

Summary

- Additive Hazards Models
- Case-Cohort Studies
- Analysis R code
  - Use Auxiliary Information to Improve Precision
  - Construct Weight
  - Handle Missing Covariates
- Biomarkers
- Open-source exercise code and dataset

Estimating Equations for Phase I Sample Estimator $\hat{\alpha}$

$$\frac{1}{N}\sum_{i=1}^{N}\Psi_{\alpha}(X_i) = 0$$



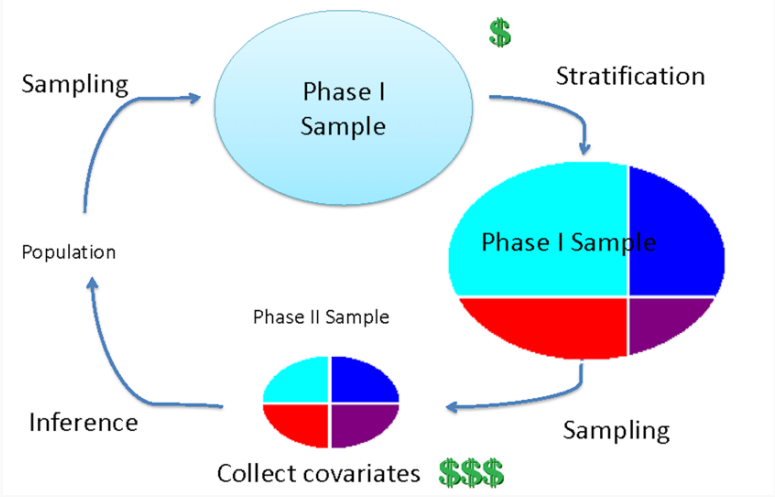
$\alpha = (\theta, \Lambda)$

N : phase I sample size

$\Psi_{\alpha}(X_i)$: a function of X derived from a model

Solution: IPW Estimating Equations for Two-phase Sampling Estimator $\hat{\theta}^*$

$$\frac{1}{N}\sum_{i=1}^{N}\frac{R_i}{\pi_0\,(V_i)}\Psi_\alpha(X_i)=0$$



$\alpha = (\theta, \Lambda)$

N : phase I sample size

R_i : phase II subsample membership

V_i : phase I variables

X_i : phase II variables

$\pi_0\,(V_i)$: phase II subsample selection probability

$\Psi_\alpha(X_i)$: a function of X derived from a model

Weight Calibration to Incorporate Auxiliary Variables

$\tilde{V}_i$: auxiliary variables (calibration variables) allowing to be a function of *any* phase I variables V_i

$$\frac{1}{N}\sum_{i=1}^{N}\frac{R_i}{\pi_0\,(V_i)}\exp(-\gamma^T\tilde{V}_i)\Psi_\alpha(X_i)=0$$

$$\frac{1}{N}\sum_{i=1}^{N}\tilde{V}_i-\frac{1}{N}\sum_{i=1}^{N}\left[\frac{R_i}{\pi_0\,(V_i)}\exp\left(-\gamma^T\tilde{V}_i\right)\tilde{V}_i\right]=0$$

Phase I observation

Phase II estimates

New Inverse Probability Weighted Estimating Equation for Calibrated Z-estimator $\hat{a}^{**}$

$$\frac{1}{N}\sum_{i=1}^{N}\frac{R_i}{\pi_0\,(V_i)}\exp(-\gamma^T\tilde{V}_i)\Psi_a(X_i)=0$$

$$\frac{1}{N}\sum_{i=1}^{N}\left[\frac{R_i}{\pi_0\,(V_i)}\exp\left(-\gamma^T\tilde{V}_i\right)\tilde{V}_i-\tilde{V}_i\right]=0$$

$\tilde{V}_i$: auxiliary variables allowing to be a function of *any* phase I variables V_i

Inverse Probability Weighted Estimating Equation

$$\mathbb{P}_N \psi_\alpha^*(X, V, R)h = \mathbb{P}_N \frac{R}{\pi_0(V)} \psi_\alpha(X)h = 0$$

α is in a Banach space, e.g., $\alpha = (\theta, \Lambda)$

Ψ_a is a map in a Banach space: $\Psi_a \in l^\infty(H)$

Given $h \in H$, $\Psi_a: h \mapsto \Psi_a h$, $\Psi_a h \in R$

Solution: A Z–estimation System

- Inverse probability weighted estimation equation (IPW-EE)*
- Calibration**
- Huber's Z-estimation#
- Modern empirical process theory##

* *Horvitz and Thompson (1952), Binder (1992)*
** *Deville and Sarndal (1992)*
\# *Huber (1967)*
\#\# *Dudley (1978), van der Vaart & Wellner (1996), van der Vaart(1998) among others*

Main Conditions

Condition 2.3.5. *the class* $\mathcal{F} \equiv \{\psi_{\alpha,h} : \|\alpha - \alpha_0\| < \delta, h \in \mathcal{H}\}$ *is P-Donsker for some* $\delta > 0$, *with finite envelope function.*

Condition 2.3.6. *as a map into* $l^\infty(\mathcal{H})$, *the map* $\alpha \mapsto P\psi_\alpha$ *is Fréchet-differentiable at a zero* α_0, *with a derivative* $\dot{\Psi}_0 : lin\mathbb{A} \mapsto l^\infty(\mathcal{H})$ *that has a continuous inverse on its range.*

Condition 2.3.7. $\|P(\psi_{\alpha,h} - \psi_{\alpha_0,h})^2\|_{\mathcal{H}} \to 0$ *as* $\alpha \to \alpha_0$.

Reference

Norm E. Breslow, **Jie Hu,** Jon A. Wellner. Z-estimation and Stratified Samples: Application to Survival Models. Lifetime Data Analysis 21, 493-516.

Jie Hu. A Z-estimation system for two-phase sampling with applications to additive hazards models and epidemiologic studies.*University of Washington ResearchWorks Archive* PhD Diss.

Jie Hu, Norman E. Breslow, Gary Chan, Couper David. Estimating the Hazard Difference from Case-Cohort Studies", European Journal of Epidemiology 36(11), 1129-1142.

Norman Breslow and **Jie Hu**. Survival Analysis of Case-Control Data: A Sample Survey Approach. Handbook of Statistical Methods for Case-Control Studies, Chapman and Hall/CRC.

Reference

Victoria Ding and **Jie Hu,** R Shiny app: Additive Hazards

Jie Hu, Fit Additive Hazards Models for Survival Analysis, CRAN - Package addhazard