Nonparametric Estimation of the Potential Impact Fraction and the Population Aggregate Fraction

Colleen Chan

Ph.D. Candidate, Dept. of Statistics and Data Science, Yale University

Collaborators: Rodrigo Zepeda-Tello[1], Dalia Camacho García Formentí[1], Frederick Cudhea[2], Rafael Meza[3], Eliane Rodrigues[4], Tonatiuh Barrientos Gutierrez[1], Donna Spiegelman[5], Xin Zhou[5]

[1]National Institute of Public Health of Mexico, [2]Tufts University Friedman School of Nutrition Science and Policy, [3]University of Michigan School of Public Health, [4]Universidad Nacional Autonoma de Mexico, [5]Yale School of Public Health

CMIPS-YCAS Seminar

October 29, 2021

Motivating Question: SSB's and type 2 diabetes

Sugar-sweetened beverages (SSB's)

- Drinks with added sugar
- The largest source of added sugar in our diets. Per-capita availability of SSB's has tripled since 1954[1]
- SSB consumption linked to increased risk of T2D, obesity[2]

Q: What fraction of type 2 diabetes cases can be attributed to SSB consumption? What if SSB consumption were entirely eliminated? What if it were halved?

[1] U.S. Department of Agriculture, U.S. Department of Health and Human Services
[2] American Heart Association

Outline

1. The PIF and PAF
    - Standard Approach
    - Mixture Approach
2. Methods
    - Empirical Method
    - Approximate Method
3. Illustrative Example
4. Simulation Studies

The PIF and PAF

- The potential impact fraction, or the attributable fraction, is the proportion of incidents attributable to a given risk factor
- It requires a relative risk function, RR that depends on exposure levels $\boldsymbol{X}$ and regression coefficients β
    - Most common form $RR(\boldsymbol{X};\beta) = \exp(\boldsymbol{X}\beta)$

Definition

The **potential impact fraction (PIF)** is defined as

$$\text{PIF} = \frac{\mathbb{E}_{\boldsymbol{X}}^{\text{obs}}\left[RR\left(\boldsymbol{X};\beta\right)\right] - \mathbb{E}_{\boldsymbol{X}}^{\text{cft}}\left[RR\left(\boldsymbol{X};\beta\right)\right]}{\mathbb{E}_{\boldsymbol{X}}^{\text{obs}}\left[RR\left(\boldsymbol{X};\beta\right)\right]}, \tag{1}$$

where $\mathbb{E}_{\boldsymbol{X}}^{\text{obs}}\left[RR(\boldsymbol{X};\beta)\right]$ represents the expected value of the relative risk under the observed exposure distribution. $=$

The PIF and PAF

- The population attributable fraction (PAF), or the attributable fraction for the population, is a specific case of the PIF when the counterfactual exposure is 0 ($\mathbb{E}^{\text{cft}}_{\boldsymbol{X}}\left[RR(\boldsymbol{X};\beta)\right] = 1$)

Definition

The **population attributable fraction (PAF)** is defined as

$$\text{PAF} = 1 - \frac{1}{\mathbb{E}^{\text{obs}}_{\boldsymbol{X}}\left[RR\big(\boldsymbol{X};\beta\big)\right]}, \tag{2}$$

where $\mathbb{E}^{\text{obs}}_{\boldsymbol{X}}\left[RR(\boldsymbol{X};\beta)\right]$ represents the expected value of the relative risk under the observed exposure distribution in a given population and $\mathbb{E}^{\text{cft}}_{\boldsymbol{X}}\left[RR(\boldsymbol{X};\beta)\right]$ is the expected value of the relative risk under a counterfactual distribution of the exposure.

Standard Approach

1. Assume a parametric distribution for exposure $\boldsymbol{X}$ (e.g. Log Normal, Weibull, Gamma)
2. Fit the parameters using method of moments estimation, matching the mean and variance of the observed exposure data
3. Estimate the PIF from Eq. 1 or the PAF from Eq. 2 using analytic or numerical integration

Issues with the standard approach:

1. PIF is undefined for heavy-tailed exposure distributions
2. PIF can be heavily biased if exposure distribution is misspecified

Standard Approach

Table 1

Relative Bias Percentage of the PAF under different distributional assumptions for the standard method.

True distribution $p_0 + (1 - p_0)f(x)$			*Distribution assumed*			
p_0	$f(x)$	true PAF	Gamma	Lognormal	Normal	Weibull
0	Gamma($k = 1.15, \theta = 0.78$)	0.2315	0	332	-37.6	-0.1
0.05		0.2225	-1.1	349.4	-41.2	-1.2
0.25		0.1843	-5	442.6	-62.3	-5.1
0.5		0.1309	-8.8	663.9	-129.8	-8.7
0.75		0.07	-11.3	1327.9	-437.9	-11
0	Weibull($k = 1.2, \lambda = 1.66$)	0.3818	0.3	161.9	-12.9	0
0.05		0.3698	-1.5	170.4	-15.2	-1.8
0.25		0.3166	-7.8	215.9	-26.4	-8
0.5		0.2359	-14.1	323.8	-55.2	-14.1
0.75		0.1338	-18.8	647.7	-186	-18.5

(Kehoe) Mixture Approach

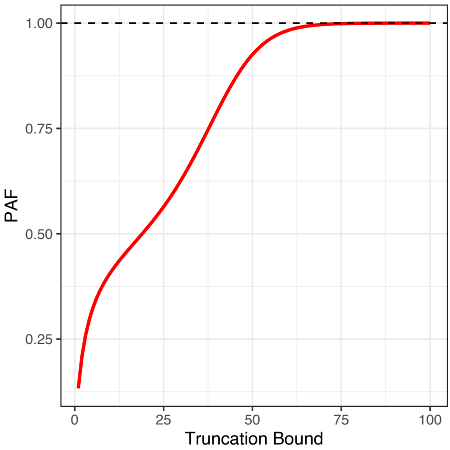
To avoid undefined PIF values, Kehoe et al. (2012) proposes:

- Truncate the assumed exposure distribution by an upper bound M
- Fit the exposure data using maximum likelihood estimation. Separate out 0 and positive values of the exposure

$$\text{PAF} = 1 - \frac{1}{p_0 RR_0 + \int_0^M RR(\boldsymbol{X};\beta) f(\boldsymbol{X}) \mathrm{d}\boldsymbol{X}}$$

Mixture Approach

PAF value now depends on truncation bound!



Population Attributable Fraction as a function of the truncation limit M considering $\boldsymbol{X}$ to be lognormally distributed with parameters $\log \mu = 0.05, \log \sigma = 0.49$ and an exponential relative risk function $RR(\boldsymbol{X}; \beta) = \exp(\beta \boldsymbol{X})$, where $\beta = log(1.3)$. Parameters are taken from the illustrative example.

Methods - Empirical Method

Let $\hat{\mu}_n^{\mathrm{obs}}(\hat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n} RR\big(\boldsymbol{X}_i; \hat{\boldsymbol{\beta}}\big)$ and $\hat{\mu}_n^{\mathrm{cft}}(\hat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n} RR\big(g(\boldsymbol{X}_i); \hat{\boldsymbol{\beta}}\big)$.
We define the empirical estimators of PAF and PIF as:

$$\widehat{\mathrm{PAF}} := 1 - \frac{1}{\hat{\mu}_n^{\mathrm{obs}}(\hat{\boldsymbol{\beta}})}, \quad \text{and} \quad \widehat{\mathrm{PIF}} := 1 - \frac{\hat{\mu}_n^{\mathrm{cft}}(\hat{\boldsymbol{\beta}})}{\hat{\mu}_n^{\mathrm{obs}}(\hat{\boldsymbol{\beta}})}.$$

Theorem

Suppose that $\hat{\boldsymbol{\beta}}$ is a consistent and asymptotically normal estimator from an independent study. That is, $\sqrt{m}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is asymptotically mean-zero multivariate normal with covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$, where m is the sample size of the independent study estimating $\boldsymbol{\beta}$. Then $\widehat{PAF}$ and $\widehat{PIF}$ converge in probability to PAF and PIF, respectively, and both $\sqrt{n}(\widehat{PAF} - PAF)$ and $\sqrt{n}(\widehat{PIF} - PIF)$ are asymptotically mean-zero multivariate normal.

Methods - Empirical Method

Sketch of proof:

1. First show asymptomic normality for $\hat{\mu}_n^{\mathrm{obs}}(\hat{\boldsymbol{\beta}})$ (and $\hat{\mu}_n^{\mathrm{cft}}(\hat{\boldsymbol{\beta}})$).
   - Note that

$$\begin{aligned}&\hat{\mu}_n^{\mathrm{obs}}(\hat{\boldsymbol{\beta}}) - \mathbb{E}(RR(X;\boldsymbol{\beta}))\\&= \left(\hat{\mu}_n^{\mathrm{obs}}(\hat{\boldsymbol{\beta}}) - \mathbb{E}(RR(X;\hat{\boldsymbol{\beta}}))\right) + \left(\mathbb{E}(RR(X;\hat{\boldsymbol{\beta}})) - \mathbb{E}(RR(X;\boldsymbol{\beta}))\right)\end{aligned}$$

   - Show that each term is asymptotically normal
   - Show that the covariance between these terms is 0
2. Derive the asymptotic normality for $\widehat{\mathsf{PIF}}$ and $\widehat{\mathsf{PAF}}$ by Delta Method

Methods - Empirical Method

We derive confidence intervals for $\widehat{\text{PIF}}$ and $\widehat{\text{PAF}}$.

$$\widehat{Var}(\hat{\mu}_n^{\text{obs}}(\hat{\boldsymbol{\beta}})) \approx \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n}(RR(X_i;\hat{\boldsymbol{\beta}}))^2 - \left(\hat{\mu}_n^{\text{obs}}(\hat{\boldsymbol{\beta}})\right)^2\right) +$$
$$\left(\frac{1}{n}\sum_{i=1}^{n}\nabla_{\boldsymbol{\beta}} RR(X_i;\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}\right)\widehat{Var}(\hat{\boldsymbol{\beta}})\left(\frac{1}{n}\sum_{i=1}^{n}\nabla_{\boldsymbol{\beta}} RR(X_i;\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}\right)^T.$$

By the delta method, the variance of $\widehat{\text{PAF}}$ can be estimated by

$$\widehat{Var}(\widehat{\text{PAF}}) \approx \frac{\widehat{Var}(\hat{\mu}_n^{\text{obs}}(\hat{\boldsymbol{\beta}}))}{(\hat{\mu}_n^{\text{obs}}(\hat{\boldsymbol{\beta}}))^4}.$$

Then the $(1-\alpha)\%$ confidence interval for $\widehat{\text{PAF}}$ is estimated as $\widehat{\text{PAF}} \pm z_{1-\frac{\alpha}{2}}\sqrt{\widehat{Var}(\widehat{\text{PAF}})}$ with $z_{1-\frac{\alpha}{2}}$ being the z-score corresponding to the $1-\frac{\alpha}{2}$ quantile of the standard normal distribution.

Methods - Approximate Method

Suppose we only had the mean and variance of the exposure.
This is often what is reported in publications, where individual-level data is not available.

To derive a point estimate using only the mean and variance of the exposure, we can use a second-order Taylor expansion.

For a twice-differentiable $h(\boldsymbol{X})$,

$$
\begin{aligned}
h(\boldsymbol{X}) &\approx h(\widehat{\boldsymbol{\mu}}) + \boldsymbol{D}h(\widehat{\boldsymbol{\mu}})(\boldsymbol{X} - \widehat{\boldsymbol{\mu}}) + \frac{1}{2}(\boldsymbol{X} - \widehat{\boldsymbol{\mu}})^T \boldsymbol{H}h(\widehat{\boldsymbol{\mu}})(\boldsymbol{X} - \widehat{\boldsymbol{\mu}}) \\
&= h(\widehat{\boldsymbol{\mu}}) + \boldsymbol{D}h(\widehat{\boldsymbol{\mu}})(\boldsymbol{X} - \widehat{\boldsymbol{\mu}}) + \frac{1}{2}tr\left[(\boldsymbol{X} - \widehat{\boldsymbol{\mu}})(\boldsymbol{X} - \widehat{\boldsymbol{\mu}})^T \boldsymbol{H}h(\widehat{\boldsymbol{\mu}})\right],
\end{aligned}
$$

where $\boldsymbol{D}h(\widehat{\boldsymbol{\mu}})$ and $\boldsymbol{H}h(\widehat{\boldsymbol{\mu}})$ are the first and second derivatives of $h(\widehat{\boldsymbol{\mu}})$ w.r.t. to $\boldsymbol{X}$

Methods - Approximate Method

Using this, we can approximate

$$
\begin{aligned}
\hat{\mu}_n^{\mathrm{obs}}(\hat{\boldsymbol{\beta}}) &:= \frac{1}{n}\sum_{i=1}^{n} RR\big(\boldsymbol{X}_i; \hat{\boldsymbol{\beta}}\big) \\
&\approx RR(\bar{\boldsymbol{X}}; \hat{\boldsymbol{\beta}}) + \frac{1}{2}\sum_{i,j} \hat{\sigma}_{i,j} \frac{\partial^2 RR\left(\boldsymbol{X}; \hat{\boldsymbol{\beta}}\right)}{\partial X_i \partial X_j}\Big|_{\boldsymbol{X}=\bar{\boldsymbol{x}}} \\
&\approx \exp(\hat{\beta}\bar{X})\left(1 + \frac{1}{2}\hat{\beta}^2\sqrt{\widehat{\mathrm{Var}}(X)}\right),
\end{aligned}
$$

leading to the following PAF estimate

$$
\widehat{\mathrm{PAF}} = 1 - \frac{1}{\exp(\hat{\beta}\bar{X})\left(1 + \frac{1}{2}\hat{\beta}^2\sqrt{\widehat{\mathrm{Var}}(X)}\right)}.
$$

Repeat for $\hat{\mu}_n^{\mathrm{cft}}(\hat{\boldsymbol{\beta}})$ for PIF estimate.

Methods - Approximate Method

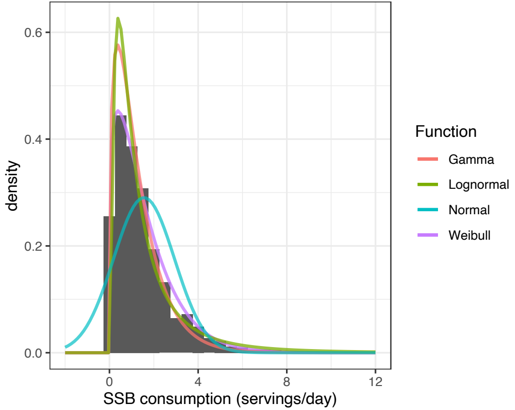
For the variance, we can apply the multivariate delta method.

Note that the PAF is a function of three components

$$\widehat{\mathsf{PAF}} = 1 - \frac{1}{\exp(\hat{\beta}\bar{X})\left(1 + \frac{1}{2}\hat{\beta}^2\sqrt{\widehat{\mathsf{Var}}(X)}\right)} = h(\bar{X}, \widehat{\mathsf{Var}}(X), \hat{\beta})$$

Illustrative Example

- Q: What proportion of type 2 diabetes cases can be attributed to sugar-sweetened beverage consumption in Mexico?
- Data from:
  - ENSANUT 2016, a national nutrition survey of the Mexican population
  - Mexican Teacher's Cohort, a longitudinal study of female Mexican teachers



Illustrative Example

	Parameters	PAF (95% CI)
Standard Gamma	$k = 1.152, \theta = 1.287$	0.2310
Mixture Gamma	$k = 1.407, \theta = 0.903$	0.3058
Mixture Gamma ($M = 12$)	$k = 1.407, \theta = 0.903$	0.3058
Standard Lognormal	$\log \mu = 0.082, \log \sigma = 0.312$	1
Mixture Lognormal	$\log \mu = 0.048, \log \sigma = 0.980$	1
Mixture Lognormal ($M = 12$)	$\log \mu = 0.048, \log \sigma = 0.980$	0.4177
Standard Normal	$\mu = 1.483, \sigma = 1.381$	1
Mixture Normal	$\mu = 1.558, \sigma = 1.374$	0.4060
Mixture Normal ($M = 12$)	$\mu = 1.558, \sigma = 1.374$	0.4060
Standard Weibull	$k = 1.075, \lambda = 1.525$	0.3772
Mixture Weibull	$k = 1.201, \lambda = 1.662$	0.3704
Mixture Weibull ($M = 12$)	$k = 1.201, \lambda = 1.662$	0.3702
Empirical	-	0.3779 (0.2655, 0.4903)
Approximate	-	0.3641 (0.2879, 0.4403)

Simulation Studies

- Define true exposure as a mixture $p_0 + (1 - p_0)f(x)$, where $f(x)$ is a known parametric distribution, truncated at $M = 12$.
  Get true PAF value.
- For each simulation $b = 1, \ldots, B$, varying N
  - Generate data from true underlying exposure distribution
  - Estimate the PAF and 95% confidence interval using the approximate and empirical methods
- Report coverage and average percent relative bias over the B simulations

Simulation Studies

				Avg % Relative Bias		Coverage	
$f(x)$	p_0	N	true PAF	empir.	approx.	empir.	approx.
Lognormal($\log\mu = 0.05, \log\sigma = 0.49$)	0.00	100	0.2778	-0.23	1.45	0.9427	0.9386
		1000		-0.21	1.47	0.9479	0.9416
		10000		-0.22	1.46	0.9484	0.9441
	0.05	100	0.2677	-0.29	1.41	0.9546	0.9514
		1000		-0.18	1.51	0.9483	0.9454
		10000		-0.22	1.47	0.9474	0.9425
	0.25	100	0.2239	-0.27	1.46	0.9504	0.9502
		1000		-0.03	1.66	0.9491	0.9466
		10000		-0.18	1.50	0.9501	0.9471
	0.50	100	0.1613	-0.18	2.27	0.9349	0.9406
		1000		0.09	2.45	0.9488	0.9451
		10000		0.27	2.61	0.9457	0.9438
	0.75	100	0.0877	0.38	6.94	0.9219	0.9244
		1000		0.20	6.74	0.9471	0.9408
		10000		0.27	6.81	0.9482	0.9408
Normal($\mu = 1.56, \sigma = 1.37$)	0.00	100	0.3937	-0.52	-0.76	0.9425	0.9453
		1000		-0.35	-0.64	0.9471	0.9475
		10000		-0.37	-0.67	0.9471	0.9482
	0.05	100	0.3815	-0.48	-1.01	0.9533	0.9565
		1000		-0.32	-0.91	0.9481	0.9491
		10000		-0.36	-0.96	0.9454	0.9473
	0.25	100	0.3275	-0.45	-2.29	0.9479	0.9481
		1000		-0.12	-2.05	0.9486	0.9494
		10000		-0.30	-2.23	0.9498	0.9479
	0.50	100	0.2451	-0.15	-3.35	0.9385	0.9283
		1000		0.02	-3.33	0.9490	0.9424
		10000		0.22	-3.17	0.9451	0.9419
	0.75	100	0.1397	0.02	-1.96	0.9178	0.9062
		1000		0.27	-1.92	0.9461	0.9405
		10000		0.39	-1.83	0.9474	0.9444

Recap

- PIF estimation requires an assumed distribution for the exposure
    - Biased when exposure distribution is misspecified and undefined when a heavily-tailed distribution is chosen
- We propose two nonparametric methods to estimate the PIF, both of which do not require making any distributional assumptions
    - Empirical method: Requires individual-level data
    - Approximate method: Requires only the mean and variance
- Conducted simulation studies of our methods
- Applied to PAF estimate of SSB consumption on type 2 diabetes incidence (≈ 0.37)

Thank you!

Back-up Slides

Methods - Approximate Method

Suppose we only had the mean $\bar{\boldsymbol{X}} = (\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k)$ and variance $\hat{\sigma}_{i,j} = Cov(X_i, X_j)$ of the exposure.

By the delta method,

$$\hat{\mu}^{\text{obs}}(\hat{\beta}) \approx RR(\bar{\boldsymbol{X}}; \hat{\beta}) + \frac{1}{2} \sum_{i,j} \hat{\sigma}_{i,j} \frac{\partial^2 RR\left(\boldsymbol{X}; \hat{\beta}\right)}{\partial X_i \partial X_j}\Big|_{\boldsymbol{X}=\bar{\boldsymbol{X}}},$$

leading to the estimator of the PAF

$$\widehat{\text{PAF}} = 1 - \frac{1}{RR(\bar{\boldsymbol{X}}; \hat{\beta}) + \frac{1}{2} \sum_{i,j} \hat{\sigma}_{i,j} \frac{\partial^2 RR\left(\boldsymbol{X}, \hat{\beta}\right)}{\partial X_i \partial X_j}\Big|_{\boldsymbol{X}=\bar{\boldsymbol{X}}}},$$

Methods - Approximate Method

Consider a general function $h(\boldsymbol{X})$, which is twice differentiable. Let

$$\boldsymbol{D}h(\boldsymbol{X}) = \frac{\partial h(\boldsymbol{X})}{\partial \boldsymbol{X}} \quad \text{and} \quad \boldsymbol{H}h(\boldsymbol{X}) = \frac{\partial^2 h(\boldsymbol{X})}{\partial \boldsymbol{X} \partial \boldsymbol{X}^T}.$$

The second-order Taylor polynomial for $h(\boldsymbol{X})$ is

$$\begin{aligned} h(\boldsymbol{X}) &\approx h(\widehat{\boldsymbol{\mu}}) + \boldsymbol{D}h(\widehat{\boldsymbol{\mu}})(\boldsymbol{X} - \widehat{\boldsymbol{\mu}}) + \frac{1}{2}(\boldsymbol{X} - \widehat{\boldsymbol{\mu}})^T \boldsymbol{H}h(\widehat{\boldsymbol{\mu}})(\boldsymbol{X} - \widehat{\boldsymbol{\mu}}) \\ &= h(\widehat{\boldsymbol{\mu}}) + \boldsymbol{D}h(\widehat{\boldsymbol{\mu}})(\boldsymbol{X} - \widehat{\boldsymbol{\mu}}) + \frac{1}{2} tr\left[(\boldsymbol{X} - \widehat{\boldsymbol{\mu}})(\boldsymbol{X} - \widehat{\boldsymbol{\mu}})^T \boldsymbol{H}h(\widehat{\boldsymbol{\mu}})\right]. \end{aligned}$$

The first and second moments of $\boldsymbol{X}$ are

$$\boldsymbol{\mu}_{\boldsymbol{X}} = \mathbb{E}(\boldsymbol{X}) \quad \text{and} \quad \boldsymbol{\Sigma}_{\boldsymbol{X}} = Var(\boldsymbol{X}),$$

and their estimates are

$$\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i) \quad \text{and} \quad \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_{\boldsymbol{X}})(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_{\boldsymbol{X}})^T.$$

Methods - Approximate Method

Applying the approximation to all subjects $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, ..., $\boldsymbol{X}_n$, we have

$$\frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{X}_i) \approx h(\widehat{\boldsymbol{\mu}}) + \frac{1}{2} tr\left[\widehat{\Sigma}_{\boldsymbol{X}} \boldsymbol{H} h(\widehat{\boldsymbol{\mu}})\right].$$

Using this, we can approximate the following scalar functions,

$$\hat{\mu}_n^{\mathrm{obs}}(\hat{\boldsymbol{\beta}}), \hat{\mu}_n^{\mathrm{cft}}(\hat{\boldsymbol{\beta}}), \frac{1}{n}\sum_{i=1}^{n}(RR(X_i;\hat{\boldsymbol{\beta}}))^2, \frac{1}{n}\sum_{i=1}^{n}(RR(g(X_i);\hat{\boldsymbol{\beta}}))^2, \frac{1}{n}\sum_{i=1}^{n} RR(X_i;\hat{\boldsymbol{\beta}})RR(g(X_i);\hat{\boldsymbol{\beta}}),$$

and the following vector functions, entry by entry,

$$\frac{1}{n}\sum_{i=1}^{n} \nabla_{\boldsymbol{\beta}} RR(X_i;\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \text{ and } \frac{1}{n}\sum_{i=1}^{n} \nabla_{\boldsymbol{\beta}} RR(g(X_i);\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}.$$

These calculations appear in the confidence intervals for $\widehat{\text{PAF}}$ and $\widehat{\text{PIF}}$.