Published in final edited form as:
Stat Med. 2015 January 30; 34(2): 297–306. doi:10.1002/sim.6327.

Optimal Allocation of Resources in a Biomarker Setting

Bernard Rosner[1,2], **Sara Hendrickson**[3], and **Walter Willett**[1,3]

[1]Channing Division of Network Medicine, Harvard Medical School, Boston, MA USA

[2]Department of Biostatistics, Harvard School of Public Health, Boston, MA USA

[3]Department of Nutrition and Epidemiology, Harvard School of Public Health, Boston, MA USA

SUMMARY

Nutrient intake is often measured with substantial error both in commonly used surrogate instruments such as a food frequency questionnaire (FFQ) as well as in gold standard type instruments such as a diet record (DR). If there is correlated error between the FFQ and DR, then standard measurement error correction methods based on regression calibration can produce biased estimates of the regression coefficient (λ) of true intake on surrogate intake. However, if a biomarker exists and the error in the biomarker is independent of the error in the FFQ and DR, then the method of triads can be used to obtain unbiased estimates of λ, provided that there is replicate biomarker data on at least a subsample of validation study subjects. Since biomarker measurements are expensive, for a fixed budget one can either use a design where a large number of subjects have 1 biomarker measure and only a small subsample is replicated, or have a smaller number of subjects and have most or all subjects validated. The purpose of this paper is to optimize the proportion of subjects with replicated biomarker measures, where optimization is with respect to minimizing the variance of $ln(\hat{\lambda})$. The methodology is illustrated using vitamin C intake data from the EPIC study where plasma vitamin C is the biomarker. In this example, the optimal validation study design is to have 21% of subjects with replicated biomarker measures.

Keywords

measurement error; biomarker; method of triads

1. INTRODUCTION

In nutritional epidemiology, the weighed diet record (DR) is considered the gold standard for assessing nutrient intake. However, it is expensive to obtain diet records and the food frequency questionnaire (FFQ) is usually used as an instrument to obtain dietary intake data from large numbers of people. It is well known that the FFQ and other dietary assessment methods have appreciable measurement error. To correct for measurement error, a validation study is often performed where both the FFQ (Z) and DR (X) are administered to the same subjects. The regression calibration factor estimated by the regression coefficient of DR on FFQ can then be used as an unbiased estimate of the regression coefficient (λ) of true dietary

[*]Correspondence to: Bernard Rosner, Channing Division of Network Medicine, Harvard Medical School, 181 Longwood Avenue, Boston, MA 02115 USA; telephone 617-525-2743; stbar@channing.harvard.edu.

intake (T) on Z, which can then be used for measurement error correction. However, this is only valid if measurement error in the DR and FFQ are uncorrelated, an assumption which may be violated. To address this issue this design is often enhanced with additional biomarker measurements (W). If the error in W is uncorrelated with the error in Z and X, then correlated error methods [1] can be used to estimate the regression calibration factor λ. The only requirement is that there be available replicate biomarker measurements on at least a subset of participants.

However, since biomarker measurements are expensive, it would be desirable to estimate the optimal proportion of subjects (θ) with replicate values of W, given a fixed total number of biomarker measures (B). The goal of this paper is to obtain a closed-form expression for var($\hat{\lambda}$) and to use it to estimate the optimal value of θ.

2. METHODS

2.1 Balanced Design

We let Z_{ij} = surrogate measure for the j^{th} replicate from the i^{th} subject, j=1, …, m_z, i=1, …, N; X_{ik} = gold standard measure for the k^{th} replicate from the i^{th} subject, k=1, …, m_x, i=1, …, N; W_{il} = biomarker for the l^{th} replicate from the i^{th} subject, l=1, …, $m_w \geq 2$, i=1, …, N.

Thus, each subject provides m_z replicates for the surrogate, m_x replicates for the gold standard and m_w replicates for the biomarker.

From Spiegelman, Zhao and Kim [1] we consider the model

$$\begin{array}{c} Z_{ij}=a+bx_i+r_i+e_{z_{ij}}, j=1,\ldots,m_z; i=1,\ldots,N \\ X_{ik}=x_i+s_i+e_{x_{ik}}, k=1,\ldots,m_x; i=1,\ldots,N \\ W_{il}=c+dx_i+e_{w_{il}}, l=1,\ldots,m_w \geq 2; i=1,\ldots,N \end{array} \quad (1)$$

where x_i = true intake for the i^{th} subject

r_i= person-specific bias in the surrogate measure $\sim N(0,\sigma_r^2)$

s_i= person-specific bias in the gold standard measure $\sim N(0,\sigma_s^2)$

$e_{z_{ij}}$, $e_{x_{ik}}$, $e_{w_{il}}$ are distributed $N(0,\sigma_{ez}^2), N(0,\sigma_{ex}^2), N(0,\sigma_{ew}^2)$ and are mutually independent of each other. $r_i \sim N(0,\sigma_r^2),\ s_i \sim N(0,\sigma_s^2) \text{cov}(r_i,s_i)=\rho_{rs}\sigma_r\sigma_s$, and r_i, s_i are mutually independent of $e_{z_{ij}}$, $e_{x_{ik}}$, $e_{w_{il}}$. Our goal is to estimate the regression calibration factor ($\lambda_{x|Z}$) = regression coefficient of x on Z. It can be shown from (1) that the MLE of $\lambda_{x|Z}$ is given by

$$\hat{\lambda}_{x|Z}=\frac{cov(Z_{ij},W_{ik})cov(X_{ij},W_{ik})}{cov(W_{il_1},W_{il_2})var(Z_{ij})} \quad (2)$$

We have found that, in simulation studies, that the distribution of $\hat{\lambda}_{x|Z}$ is generally skewed, while the distribution of $ln(\hat{\lambda}_{x|Z})$ is approximately normal. Hence, two-sided 100% × (1-α) confidence limits for $\lambda_{x|Z}$ are obtained from [exp(c_1), exp(c_2)], where

$(c_1, c_2) = \ln(\hat{\lambda}_{x|Z}) \pm z_{1-\alpha/2}\sqrt{var[\ln(\hat{\lambda}_{x|Z})]}$ and z_p = pth percentile of a N(0,1) distribution. It remains to derive an analytic expression for var[ln($\hat{\lambda}_{x|Z}$)]. For this purpose, we take the natural log of each side of equation 2 and obtain:

$$\begin{aligned} \ln(\hat{\lambda}_{x|Z}) = &\ln[cov(Z_{ij}, W_{ik})] + \ln[cov(X_{ij}, W_{ik})] \\ &-\ln[cov(W_{il_1}, W_{il_2})] - \ln[var(Z_{ij})] \\ \equiv\ & A + B - C - D \end{aligned} \qquad (3)$$

Thus,

$$\begin{aligned} var[\ln(\hat{\lambda}_{x|Z})] = var(A) + var(B) + var(C) + var(D) + 2cov(A, B) \\ -2cov(A, C) - 2cov(A, D) - 2cov(B, C) - 2cov(B, D) + 2cov(C, D) \end{aligned} \qquad (4)$$

We derive *var*(*A*). The other components can be derived in a similar manner. For notational purposes, it will be useful to introduce the notation:

μ_{abc} =E[$(Z_{ij} - Z)^{\bar{a}}(X_{ik} - X)^{\bar{b}}(W_{il} - W)^{\bar{c}}$] which we estimate by

$$\hat{\mu}_{abc} = \sum_{i=1}^{N}\sum_{j=1}^{m_z}\sum_{k=1}^{m_x}\sum_{l=1}^{m_w}(Z_{ij} - \overline{Z})^a (X_{ik} - \overline{X})^b (W_{il} - \overline{W})^c / (N\, m_z^{a^*} m_x^{b^*} m_w^{c^*}) \qquad (5)$$

where a^* = 1 if $a \ge 1$,=0 else, b^* =1 if $b \ge 1$,=0 else and c^* =1 if $c \ge 1$,=0 else. Using the delta method, we have that

$$var(A) = var(\hat{\mu}_{101})/(\hat{\mu}_{101})^2 \qquad (6)$$

Furthermore,

$$\begin{aligned} var(\hat{\mu}_{101}) &= var[\textstyle\sum_{i=1}^{N}\sum_{j=1}^{m_z}\sum_{k=1}^{m_w}(Z_{ij} - \overline{Z})(W_{ik} - \overline{W})/(Nm_z m_w)] \\ &= var\left[\sum_{j=1}^{m_z}\sum_{k=1}^{m_w}(Z_{ij} - \overline{Z})(W_{ik} - \overline{W})/(Nm_z^2 m_w^2)\right] \\ &= \left\{ \begin{array}{c} var[(Z_{ij} - \overline{Z})(W_{ik} - \overline{W})] + (m_w - 1)cov[(Z_{ij} - \overline{Z})(W_{ik_1} - \overline{W}), (Z_{ij} - \overline{Z})(W_{ik_2} - \overline{W})] \\ +(m_z - 1)cov[(Z_{ij_1} - \overline{Z})(W_{ik} - \overline{W}), (Z_{ij_2} - \overline{Z})(W_{ik} - \overline{W})] \\ +(m_z - 1)(m_w - 1)cov[(Z_{ij_1} - \overline{Z})(W_{ik_1} - \overline{W}), (Z_{ij_2} - \overline{Z})(W_{ik_2} - \overline{W})]/(Nm_z m_w) \end{array} \right\} \end{aligned} \qquad (7)$$

where $k_1 \ne k_2$ and $j_1 \ne j_2$. We can write

$$\begin{aligned} var[(Z_{ij} - \overline{Z})(W_{ik} - \overline{W})] = \mathrm{E}\left[(Z_{ij} - \overline{Z})^2 (W_{ik} - \overline{W})^2\right] - \mathrm{E}^2[(Z_{ij} - \overline{Z})(W_{ik} - \overline{W})] \\ = \hat{\mu}_{202} - \hat{\mu}_{101}^2 \end{aligned} \qquad (8)$$

Similarly, we can write

$$\begin{gathered}cov[(Z_{ij}-\overline{Z})(W_{ik_1}-\overline{W}),(Z_{ij}-\overline{Z})(W_{ik_2}-\overline{W})]\\ =\mathrm{E}\left[(Z_{ij}-\overline{Z})^2(W_{ik_1}-\overline{W})(W_{ik_2}-\overline{W})\right]-\mathrm{E}^2[(Z_{ij}-\overline{Z})(W_{ik}-\overline{W})]\end{gathered}$$

In general, we introduce the notation

$$\hat{\mu}_{a_1a_2\ldots a_r,b_1b_2\ldots b_s,c_1c_2\ldots c_t}=\mathrm{E}\left[\prod_{f=1}^{r}\left(Z_{ij_f}-\overline{Z}\right)^{\alpha_f}\prod_{g=1}^{s}\left(X_{ik_g}-\overline{X}\right)^{b_g}\prod_{h=1}^{t}(W_{il_h}-\overline{W})^{c_h}\right]$$

where $j_1 \neq j_2 \neq \cdots \neq j_r$, $k_1 \neq k_2 \neq \cdots \neq k_s$, and $l_1 \neq l_2 \neq \cdots \neq l_t$.

Thus, we have:

$$cov[(Z_{ij}-\overline{Z})(W_{ik_1}-\overline{W}),(Z_{ij}-\overline{Z})(W_{ik_2}-\overline{W})]=\hat{\mu}_{2,0,11}-\hat{\mu}^2_{101} \quad (9)$$

Similarly,

$$cov[(Z_{ij_1}-\overline{Z})(W_{ik}-\overline{W}),(Z_{ij_2}-\overline{Z})(W_{ik}-\overline{W})]=\hat{\mu}_{11,0,2}-\hat{\mu}^2_{101}$$

and

$$\begin{gathered}cov[(Z_{ij_1}-\overline{Z})(W_{ik_1}-\overline{W}),(Z_{ij_2}-\overline{Z})(W_{ik_2}-\overline{W})]\\ =\mathrm{E}[(Z_{ij_1}-\overline{Z})(Z_{ij_2}-\overline{Z})(W_{ik_1}-\overline{W})(W_{ik_2}-\overline{W})]-\hat{\mu}^2_{101}=\hat{\mu}_{11,0,11}-\hat{\mu}^2_{101}\end{gathered} \quad (10)$$

Upon combining equations 6–10, we obtain

$$var(A)=\frac{1}{\hat{\mu}^2_{101}Nm_zm_w}\left[\begin{gathered}\hat{\mu}_{202}+(m_w-1)\hat{\mu}_{2,0,11}+(m_z-1)\hat{\mu}_{11,0,2}+(m_w-1)(m_z-1)\hat{\mu}_{11,0,11}\\ -m_wm_z\hat{\mu}^2_{101}\end{gathered}\right] \quad (11)$$

The other components in equation 4 are obtained similarly and are provided in Web Appendix A.

Upon combining equations A1–A10, we obtain $var[ln(\hat{\lambda}_{x|Z})]$ in equation 4.

To obtain confidence limits for $\lambda_{x|Z}$ we assume asymptotic normality of $\ln(\hat{\lambda}_{x|Z})$ whereby a two-sided 100% × (1-α) CI for $\lambda_{x|Z}$ is given by [exp(c_1), exp (c_2)], where

$$(c_1,c_2)=\ln(\hat{\lambda}_{x|Z})\pm z_{1-\alpha/2}\sqrt{var[\ln(\hat{\lambda}_{x|Z})]} \quad (12)$$

and $z_{1-\alpha/2}$= upper $\alpha/2$ percentile of a N(0,1) distribution.

2.2 Unbalanced Design

We now consider the unbalanced design situation. In this case, we assume all subjects have the same number of replicates for the surrogate dietary instrument (e.g., FFQ) and the gold standard dietary instrument (e.g., DR) denoted by m_z and m_x, respectively. However, since biomarker measurements are the most expensive, we assume that n_g of the subjects have g biomarker measurements, where $g = 1,2$ and $n_1 + n_2 = N$. Also, let b_i = the number of replicate biomarker measurements for the i^{th} subject and let $M=\sum_{i=1}^{N} b_i=2n_2+n_1$. Finally, let θ= proportion of biomarker measurements that are replicated $=2n_2/M$ where $0 \le \theta \le 1$. We assume that M is fixed due to budgetary constraints and we wish to determine the value of θ that minimizes $var[ln(\hat{\lambda}_{x|Z})]$ in equation 4.

We will derive $var(A)$ in the unbalanced case and present the results for the other components of equation 4 in Appendix B. In the unbalanced case, we estimate μ_{abc} by

$$\hat{\mu}_{abc}=\sum_{i=1}^{N}\sum_{j=1}^{m_z}\sum_{k=1}^{m_x}\sum_{l=1}^{b_i}(Z_{ij}-\overline{Z})^a(X_{ik}-\overline{X})^b(W_{il}-\overline{W})^c/m_z^{a^*}m_x^{b^*}M \quad (13)$$

where a^* and b^* are defined in equation 5.

We have

$$var(A)=var(\hat{\mu}_{101})/\hat{\mu}_{101}^2 \quad (14)$$

where μ_{101} is estimated using equation 13.

We have:

$$var(\hat{\mu}_{101})=\frac{1}{m_z^2M^2}\sum_{i=1}^{N} var[\sum_{j=1}^{m_z}\sum_{l=1}^{b_i}(Z_{ij}-\overline{Z})(W_{il}-\overline{W})] \quad (15)$$

Furthermore,

$$\begin{gathered}var[\sum_{j=1}^{m_z}\sum_{l=1}^{b_i}(Z_{ij}-\overline{Z})(W_{il}-\overline{W})]=m_zb_i\, var[(Z_{ij}-\overline{Z})(W_{il}-\overline{W})]\\ +m_zb_i(b_i-1)cov[(Z_{ij}-\overline{Z})(W_{il_1}-\overline{W}),(Z_{ij}-\overline{Z})(W_{il_2}-\overline{W})]\\ +m_z(m_z-1)b_i\, cov[(Z_{ij_1}-\overline{Z})(W_{il}-\overline{W}),(Z_{ij_2}-\overline{Z})(W_{il}-\overline{W})]\\ +m_z(m_z-1)b_i(b_i-1)cov[(Z_{ij_1}-\overline{Z})(W_{il_1}-\overline{W}),(Z_{ij_2}-\overline{Z})(W_{il_2}-\overline{W})]\\ =m_zb_i(\hat{\mu}_{202}-\hat{\mu}_{101}^2)+m_zb_i(b_i-1)(\hat{\mu}_{2,0,11}-\hat{\mu}_{101}^2)+m_z(m_z-1)b_i(\hat{\mu}_{11,0,2}-\hat{\mu}_{101}^2)\\ +m_z(m_z-1)b_i(b_i-1)(\hat{\mu}_{11,0,11}-\hat{\mu}_{101}^2)\end{gathered} \quad (16)$$

If we denote $\sum_{i=1}^{N} b_i^2$ by $M^{(2)}$ and combine equations 14, 15 and 16, we obtain

$$var(A)=\frac{1}{m_z M^2 \hat{\mu}_{101}^2}\{M(\hat{\mu}_{202}-\hat{\mu}_{101}^2)+[M^{(2)}-M]\,(\hat{\mu}_{2,0,11}-\hat{\mu}_{101}^2)+(m_z-1)M(\hat{\mu}_{11,0,2}-\hat{\mu}_{101}^2)+ (m_z-1)\,[M^{(2)}-M]\,(\hat{\mu}_{11,0,11}-\hat{\mu}_{101}^2)\} \quad (17)$$

Note, if there are a total of N subjects of whom n_1 have one replicate and n_2 have two replicates, then $M = 2n_2 + n_1$, $M^{(2)} = 4n_2 + n_1$, and $M^{(2)} - M = 2n_2$. In this case, equation 17 reduces to:

$$var(A)_{unbalanced}=\frac{1}{m_z(2n_2+n_1)^2\hat{\mu}_{101}^2}\{(2n_2+n_1)(\hat{\mu}_{202}-\hat{\mu}_{101}^2)+2n_2(\hat{\mu}_{2,0,11}-\hat{\mu}_{101}^2) +(m_z-1)(2n_2+n_1)(\hat{\mu}_{11,0,2}-\hat{\mu}_{101}^2)+(m_z-1)2n_2(\hat{\mu}_{11,0,11}-\hat{\mu}_{101}^2)\} \quad (18)$$

Derivation of the other components of equation 4 under an unbalanced design are obtained similarly and are provided in Web Appendix B.

Finally, a large sample 100% × (1-α) CI for $\lambda_{x|Z}$ is given by [exp(c_1), exp(c_2)] where

$$(c_1,c_2)=\ln(\hat{\lambda}_{x|Z}) \pm z_{1-\alpha/2}\sqrt{var[\ln(\hat{\lambda}_{x|Z})]}$$

2.3 Optimization

We wish to minimize $var[ln(\hat{\lambda}_{x|Z})]$ in equation 4 in the setting where b_i= 1 or 2. We can re-express equation B.1 in Web Appendix B as a function of θ as follows:

$$var(A)=f_{1A}+f_{2A}\theta \quad (19)$$

where

$$\theta=2n_2/(n_1+2n_2)$$
$$f_{1,A}=\frac{\hat{\mu}_{202}-\hat{\mu}_{101}^2+(m_z-1)(\hat{\mu}_{11,0,2}-\hat{\mu}_{101}^2)}{m_z\hat{\mu}_{101}^2 M}$$
$$f_{2,A}=\frac{\hat{\mu}_{2,0,11}-\hat{\mu}_{101}^2+(m_z-1)(\hat{\mu}_{11,0,11}-\hat{\mu}_{101}^2)}{m_z\hat{\mu}_{101}^2 M}$$

Similarly,

$$var(B)=f_{1B}+\theta f_{2B} \quad (20)$$

$$var(C)=f_C/\theta \quad (21)$$

$$var(D)=f_D/(2-\theta) \quad (22)$$

$$cov(A,B)=f_{1,AB}+\theta f_{2,AB} \quad (23)$$

$$cov(A,C)=f_{AC} \quad (24)$$

$$cov(A,D)=f_{AD}/(2-\theta) \quad (25)$$

$$cov(B,C)=f_{BC} \quad (26)$$

$$cov(B,D)=f_{BD}/(2-\theta) \quad (27)$$

$$cov(C,D)=f_{CD}/(2-\theta) \quad (28)$$

The expressions for f_{1A}, …, f_{CD} are given in Appendix C.

Note that in general if there is positive correlation among replicate Z, X and W values, then it can be shown that $f_{2A} > 0$, $f_{2B} > 0$, $f_C > 0$, $f_D > 0$, $f_{2,AB} > 0$, $f_{AD} > 0$, $f_{BD} > 0$, and $f_{CD} > 0$. Hence, *var* (*A*), *var* (*B*), *var* (*D*), *cov*(*A*, *B*), *cov*(*A*, *D*), *cov*(*B*, *D*) and *cov*(*C*, *D*) are minimized if $\theta = 0$, i.e., all subjects have only one biomarker measurement, since this will maximize the number of subjects. Conversely, *var*(*C*) is minimized if $\theta = 1$; where all subjects have two biomarker measurements.

Assume all subjects have either one or two biomarker measurements. If we combine equations 4 and 19–28, we obtain:

$$var[ln(\hat{\lambda}_{x|Z})]=C_0+C_1\theta+C_2/\theta+C_3/(2-\theta) \equiv V(\theta) \quad (29)$$

where

$$\begin{gathered} C_0=f_{1A}+f_{1B}+2f_{1,AB}-2f_{AC}-2f_{BC} \\ C_1=f_{2A}+f_{2B}+2f_{2,AB} \\ C_2=f_C \\ C_3=f_D-2f_{AD}-2f_{BD}+2f_{CD} \end{gathered}$$

If we differentiate $V(\theta)$ with respect to θ in equation 29 and collect terms, we obtain the 4th degree polynomial equation as follows:

$$\theta^4-4\theta^3+d_1\theta^2+d_2\theta-d_2=0 \quad (30)$$

where

$$d_1=4-\frac{(C_2-C_3)}{C_1}, d_2=\frac{4C_2}{C_1}$$

Although it is possible to obtain an exact solution to this equation, it is simpler to use a polynomial equation solver (e.g., the POLYROOT function of SAS) to determine the solution that satisfies $0 < \theta < 1$.

3. SIMULATION STUDY

We simulated data from a hypothetical dataset with a similar correlation structure as in our example with $(Z_1, Z_2, X_1, X_2, W_1, W_2) \sim N(\underline{\mu}, \Sigma)$ where $\underline{\mu} = (100, 100, 100, 100, 50, 50)$ and

$$\sum = \begin{pmatrix} 400 & 232 & 208 & 172 & 48 & 48 \\ 232 & 400 & 172 & 208 & 48 & 48 \\ 208 & 172 & 400 & 240 & 64 & 64 \\ 172 & 208 & 240 & 400 & 64 & 64 \\ 48 & 48 & 64 & 64 & 100 & 40 \\ 48 & 48 & 64 & 64 & 40 & 100 \end{pmatrix}$$

We then estimated $ln(\lambda_{x|Z})$ in equation 3, its variance in equation 4 and a 95% CI for $\lambda_{x|Z}$ in equation 12 from 4,000 simulated samples. The results are given in Table 1. We see that there is good agreement between the mean theoretical variances and covariances considered in equation 4 and derived in Appendix B and the corresponding empirical variances and covariances obtained from the 4,000 simulated samples. Also, the overall estimate of $\lambda_{x|Z}$ has little bias and the estimated 95% confidence intervals have approximately (94.1%) coverage.

4. EXAMPLE

We analyzed data from the EPIC-Norfolk study [2]. Individuals were seen at a baseline visit and at a 4-year follow-up visit as part of the study. At both baseline and follow-up, a food frequency questionnaire (FFQ) and a 1-week diet record (DR) were obtained. In addition, a blood sample was obtained at both the baseline and 4-year follow-up visit. In this example, we focus on dietary vitamin C and assess the regression coefficient of true dietary vitamin C intake (x_i in equation 1) on FFQ vitamin C intake ($\boldsymbol{Z}_{ij}$ in equation 1) which is given by $\widehat{\lambda_{x|Z}}$ in equation 2 using plasma vitamin C as a biomarker. We refer to $\widehat{\lambda_{x|Z}}$ as the estimated regression calibration factor. For this example, we assume that true dietary intake has not changed over four years, but allow for the possibility of correlated error between FFQ and DR intake (ρ_{rs} in equation 1). We also assume that there is no systematic error in the biomarker and that the random error in FFQ intake, DR intake and plasma vitamin C are uncorrelated. The marginal and joint distribution of FFQ intake ($\boldsymbol{Z}_{ij}$), DR intake (X_{ij}) and plasma vitamin C (W_{ij}) are given in Table II. There is moderate correlation between dietary vitamin C (Z, X) and plasma vitamin C (W) which are similar for the FFQ and DR when the intake assessments at year 4 are compared with the biomarker values at baseline (which provides the most appropriate assessment of their relative measurement of long-term intake). For the purpose of better approximating a normal distribution, the log transform was used for each of dietary vitamin C from FFQ ($\boldsymbol{Z}_{ij}$) and DR (X_{ik}) in subsequent analyses.

In Table III we provide the point estimate and 95% CI for $\lambda_{x|Z}$ as well as the individual components used in equation 12. We see that the estimated regression calibration factor ($\lambda_{x|Z}$) is 0.308 with 95% confidence limits from 0.201 to 0.471. The point estimate implies that there is substantial measurement error in the assessment of dietary vitamin C. For example if the estimated hazard ratio based on observed vitamin C is 1.2 then the deattenuated estimate would be $1.2^{1/0.308} = 1.8$, indicating substantial deattenuation. The degree of measurement error in the FFQ will vary depending on the nutrients/foods being considered. In general, beverage intake has less measurement error, while food intake can have considerable measurement error. Dietary vitamin C is derived mainly from fruits and vegetables which have moderate measurement error.

5. OPTIMIZATION

We also used the EPIC data to estimate the optimal proportion of replicated biomarker measurements based on equations 29 and 30. The results are presented in Table IV. The estimated parameters (C_1, C_2, C_3), (d_1, d_2) in equations 29 and 30 are given in the left side of the table. The solution using the POLYROOT function of SAS was $\hat{\theta} = 0.349$ = the optimal proportion of replicated biomarker measurements (i.e., $2n_2/(n_1 + 2n_2)$). It follows directly that the optimal estimate of $n_2/n_1 = 0.349/[2(0.651)] = 0.268$ or equivalently $n_2/(n_1 + n_2) = 0.268/1.268 = 0.211$. Thus, the optimal design (i.e., min $var[\ln(\hat{\lambda}_{x|Z})]$) is for approximately 21% of the sample to have replicated biomarker measurements given a fixed total of M biomarker measurements. To assess the sensitivity of $var[\ln(\hat{\lambda}_{x|Z})]$ to variation in θ we computed $var[\ln(\hat{\lambda}_{x|Z})]$ for different values of θ. The results are given in the right hand side of Table IV and are plotted in Figure 1. We see that the variance function is fairly flat between $\hat{\theta} = 0.2 – 0.5$ corresponding to a proportion of subjects with replicated biomarkers of 0.14 to 0.33. However, the variance increases moderately outside these limits.

6. DISCUSSION

Correlated error between gold standard dietary measures such as a diet record and surrogate measures such as a food frequency questionnaire can bias standard techniques for correcting for measurement error such as regression calibration. The method of triads using a biomarker in addition to the above dietary instruments is an effective method for eliminating this bias. However, it requires replicate measurements on the biomarker for at least a subset of study participants [1]. In the current paper, we derive a closed form expression for the variance estimate of the Spiegelman, Zhao and Kim estimator of the regression calibration factor ($\lambda_{x|Z}$) and associated 95% confidence limits for both balanced (same number of biomarker replicates per subject) and unbalanced (different number of biomarker replicates per subject) designs.

Ideally, all subjects in a validation study would have replicated biomarker measurements; however, these measures are usually expensive. Thus, in this paper, we derive an expression for the optimal proportion of validation study subjects with replicated biomarker measures given a fixed total number of biomarker measures (M), where optimality is defined as minimizing $var[\ln(\hat{\lambda}_{x|Z})]$. In the EPIC example, this was about 21%, but would be expected to vary for other biomarkers or in other studies.

The algorithms used to derive $var[\ln(\hat{\lambda}_{x|Z})]$ and associated confidence limits and the optimal design formulas in equations 29 and 30 are available in the form of SAS macros from the authors upon request.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We acknowledge the support of R01 CA50597, R01 CA077398 and U54 CA155626 from the National Institutes of Health in performing this work. We also acknowledge programming support of Rong Chen and Marion McPhee. Sara Hendrickson was supported in part by training grants a 5 T32 CA09001 and R25 CA098566.

References

1. Spiegelman D, Zhao B, Kim J. Correlated errors in biased surrogates: study designs and methods for measurement error correction. Statistics in Medicine. 2005; 24(11):1657–82. [PubMed: 15736283]
2. Rosner B, Michels KB, Chen YH, Day NE. Measurement error correction for nutritional exposures with correlated measurement error: use of the method of triads in a longitudinal setting. Statistics in Medicine. 2008; 27(18):3466–89.10.1002/sim.3238 [PubMed: 18416440]

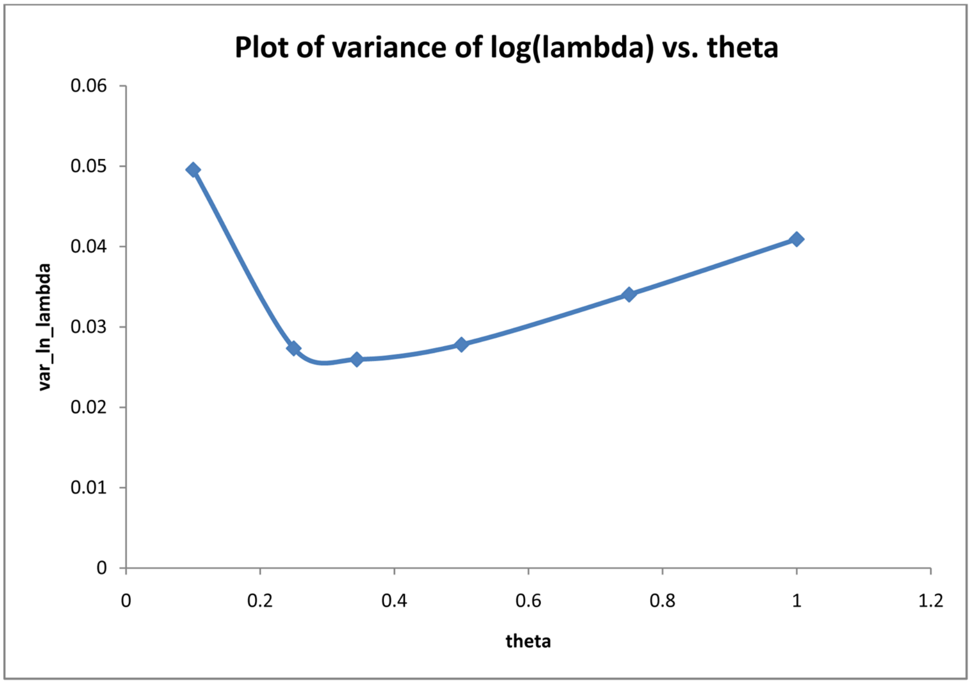
Plot of variance of log(lambda) vs. theta
var_ln_lambda
0.06
0.05
0.04
0.03
0.02
0.01
0
0
0.2
0.4
0.6
0.8
1
1.2
theta

Figure 1.

Table I

Simulation Study Results, 4000 replications

Component	Theoretical value[*]	Empirical estimate	Coverage probability
var(A)	0.0351	0.0389	
var(B)	0.0204	0.0233	
var(C)	0.0234	0.0234	
var(D)	0.0041	0.0041	
cov(A,B)	0.0169	0.0181	
cov(A,C)	0.0109	0.0116	
cov(A,D)	0.0048	0.0049	
cov(B,C)	0.0108	0.0116	
cov(B,D)	0.0022	0.0023	
cov(C,D)	0.0009	0.0010	
cov(Z_{ij}, W_{ik})	48.0	48.0	
cov(X_{ij}, W_{ik})	64.0	63.9	
cov(W_{il_1}, W_{il_2})	40.0	39.8	
var(Z_{ij})	400.0	399.6	
var[$ln(\hat{\lambda}_{x\|Z})$]	0.0609	0.0672	
$\hat{\lambda}_{x\|Z}$	0.192	0.194[**]	0.941

[*] Based on Web Appendix B

[**] median

Computer program :/proj/stross/stros0c/measurmentErrBio/Undesignx4000a.sas 09/27/13
:/proj/stross/stros0c/measurmentErrBio/all_new2.txt 09/27/13

Table II

Marginal and Joint Distribution of FFQ vitamin C, DR vitamin C and plasma vitamin C in the EPIC-Norfolk study

			correlation matrix					
variable	**mean**	**sd**	Z_{i1}	Z_{i2}	X_{i1}	X_{i2}	W_{i1}	W_{i2}
Z^*_{i1}	134.4	54.5	1.0	0.60	0.47	0.42	0.25	0.18
Z^*_{i2}	135.6	58.7		1.0	0.45	0.57	0.25	0.27
$X^\dagger_{i1}$	90.5	50.1			1.0	0.59	0.40	0.23
$X^\dagger_{i2}$	94.6	52.0				1.0	0.28	0.34
$W^\ddagger_{i1}$	57.7	21.2					1.0	0.43
$W^\ddagger_{i2}$	64.8	23.2						1.0

[*] Z_{i1}, Z_{i2} = baseline and 4-year calorie-adjusted FFQ vitamin C intake (mg/day)

[†] X_{i1}, X_{i2} = baseline and 4-year calorie-adjusted DR vitamin C intake (mg/day)

[‡] W_{i1}, W_{i2} = baseline and 4-year plasma vitamin C intake (μmol/L)

Computer program: :/proj/stross/stros0a/example_usevitc.sas 09/19/13

Table III

Estimation of Regression Calibration factor in EPIC data example, n=323

A[*]	2.998
B[*]	5.003
C[*]	263.319
D[*]	0.1852
var(A)	0.034
var(B)	0.019
var(C)	0.019
var(D)	0.007
cov(A,B)	0.018
cov(A,C)	0.013
cov(A,D)	0.009
cov(B,C)	0.011
cov(B,D)	0.004
cov(C,D)	0.003
$\widehat{\lambda_{x\|Z}}$	0.308
$log(\widehat{\lambda_{x\|Z}})$	−1.179
$var[log(\widehat{\lambda_{x\|Z}})]$	0.0473
95% CI for $\lambda_{x\|Z}$	(0.201,0.471)

[*] $A = cov(Z_{ij}, W_{ik})$; $B = cov(X_{ij}, W_{ik})$; $C = cov(W_{il_1}, W_{il_2})$; $D = var(Z_{ij})$

Computer run: :/proj/stross/stros0c/measurmentErrBio/example1/example_usevitc.sas 6/4/12

Table IV

Results of Optimization Procedure based on EPIC dataset

Parameter	**Value**	θ	$var[ln(\hat{\lambda}_{x\|Z})]$	$n_2/(n_1 + n_2)$
C_1	0.04120	0.10	0.0496	0.053
C_2	0.00472	0.25	0.0273	0.143
C_3	−0.00664	0.349	0.0259	0.208
d_1	3.72420	0.500	0.0278	0.333
d_2	0.45839	0.75	0.0340	0.600
$\hat{\theta}$	0.349	0.90	0.0382	0.818

Computer program:
:/proj/stross/stros0c/measurmentErrBio/example1/example2_usevitc.sas 9/30/13
:/proj/stross/stros0c/measurmentErrBio/example1/test_getLambda.sas 9/30/13