Published in final edited form as:

Stat Med. 2012 October 15; 31(23): 2713–2732. doi:10.1002/sim.5435.

Regression Calibration with More Surrogates than Mismeasured Variables

Victor Kipnis^{a,*†}, Douglas Midthune^a, Laurence S. Freedman^b, and Raymond J. Carroll^c

^aBiometry Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, MD, U.S.A

^bGertner Institute for Epidemiology and Health Policy Research, Sheba Medical Center, Tel Hashomer, Israel

^cDepartment of Statistics, Texas A&M University, College Station, TX, U.S.A

Abstract

In a recent paper, Weller, Milton, Eisen, and Spiegelman discuss fitting logistic regression models when a scalar main explanatory variable is measured with error by several surrogates, i.e., a situation with more surrogates than variables measured with error. Using a regression calibration approximate model as if it were exact, they compare two methods of adjusting for measurement error. One is the standard regression calibration approach consisting of substituting an estimated conditional expectation of the true covariate given observed data in the logistic regression. The other is a novel two-stage approach when the logistic regression is fitted to multiple surrogates, and then a linear combination of estimated slopes is formed as the estimate of interest. Applying estimated asymptotic variances for both methods in a single data set with some sensitivity analysis, the authors assert superiority of their two-stage approach. We investigate this claim in some detail. A troubling aspect of the proposed two-stage method is that, unlike standard regression calibration and a natural form of maximum likelihood, the resulting estimates are not invariant to reparameterization of nuisance parameters in the model. We show, however, that, under the regression calibration approximation, the two-stage method is asymptotically equivalent to a maximum likelihood formulation, and is therefore in theory superior to standard regression calibration. However, our extensive finite-sample simulations in the practically important parameter space where the regression calibration model provides a good approximation failed to uncover such superiority of the two-stage method. We also discuss extensions to different data structures.

Keywords

attenuation; logistic regression; measurement error; reparameterization

1. Introduction

1.1. Basic problem

Weller *et al.* [1] consider a problem where a binary variable Y is to be regressed on a scalar main explanatory variable X and a vector covariate $\mathbf{Z}$ via logistic regression, but that X is subject to measurement error. Instead of observing X , a vector covariate $\mathbf{W}$ is observed, so

*Correspondence to: Victor Kipnis, Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Executive Plaza North, Room 3124, 6130 Executive Blvd., MSC 7354, Bethesda, MD 20892-7354, U.S.A.

†kipnisv@mail.nih.gov, Tel: 301-496-7464, FAX: 301-402-0816

that there are more surrogates, $\mathbf{W}$, than there are variables measured with error, the scalar X . In place of standard regression calibration as described by Carroll and Stefanski [2] and Carroll *et al.* [3], they describe a novel extension of the two-stage linear approximation approach by Rosner *et al.* [4] to estimating the slope for X . They compare their new approach with standard regression calibration by estimating the corresponding approximate asymptotic variances based on a single data set and some sensitivity analysis and conclude that the two-stage approach provides a more efficient estimator than standard regression calibration.

The purpose of this paper is to place these methods in a more general context than just logistic regression, and to replace the somewhat informal calculations of Weller *et al.* with a more formal theoretical treatment. There are four important aspects of this treatment.

- Weller *et al.* work with an approximate model, namely the regression calibration approximation to the conditional distribution of Y given the observed covariates $(\mathbf{W}, \mathbf{Z})$ and then do all their theoretical calculations as if this approximate model were true. This is a fairly common strategy when dealing with measurement error modeling and regression calibration, see for example Thurston *et al.* [5]. In the parameter subspace where such an approximation is close to the correct model, this allows a formal theoretical analysis, which we provide.
- In this problem, the primary interest is in a single parameter, the regression slope for the main explanatory variable. All other parameters are nuisance parameters. We can thus write the likelihood function of the data as $L(\boldsymbol{\zeta}, \beta_1)$, where $\boldsymbol{\zeta}$ is the vector of nuisance parameters and β_1 is the parameter of interest. In all such likelihood problems, the maximum likelihood estimator of β_1 is unchanged when a 1–1 reparameterization of $\boldsymbol{\zeta}$ is done: the same is true for standard regression calibration. Unfortunately, the suggested two-stage estimator is not invariant to such 1–1 reparameterizations of the nuisance parameters, and can be made to vary quite widely in a given finite set of data. We suggest a modified two-stage estimator which, in our extensive simulations, provides better finite sample results than the original method.
- We show that if the regression calibration approximation were exact, the two-stage estimator would have the same asymptotic variance as the maximum likelihood estimator. Extensive numerical work shows that, asymptotically, the maximum likelihood estimator and modified two-stage estimator are typically only slightly less variable than the one based on standard regression calibration, at least in the practically important parameter space where the regression calibration model provides a good approximation.
- As noted above, the theoretical calculations in Weller *et al.* and in the present work are based on an approximate model. It is well known that this approximation can be a very poor one if the main explanatory variable has a relatively large effect on the outcome and/or measurement error is relatively large. In such cases, as is well demonstrated by the results in Weller *et al.*, the model based on the regression calibration assumption is badly misspecified leading to severely inconsistent estimators with very different asymptotic and finite-sample properties. Comparing asymptotic variances of estimators derived under the regression calibration assumption may be irrelevant in this situation. However, these are exactly the cases where Weller *et al.* emphasize the superiority of their two stage method over the standard one, disregarding involved biases and finite-sample properties. We therefore provide finite-sample simulation results which suggest that standard regression calibration is, in most cases, at least as good as the two-stage method in terms of bias and mean squared error.

1.2. Basic model assumptions

The setup of the problem follows Weller *et al.* We have primary study data consisting of $(Y_{ip}, \mathbf{W}_{ip}, \mathbf{Z}_{ip})$ for $i = 1, \dots, n_p$. We also have data on an external validation study $(\mathbf{W}_{iv}, X_{iv}, \mathbf{Z}_{iv})$ for $i = 1, \dots, n_v$, where X is scalar. There are two parts to the model. The first is the relationship between X and $(\mathbf{W}, \mathbf{Z})$. It is assumed that the regression of X on $(\mathbf{W}, \mathbf{Z})$ is linear and homoscedastic in both the primary and the validation data with identical parameters. In symbols, $X_{iv} = \gamma_0 + \mathbf{W}_{iv}^T \boldsymbol{\gamma}_1 + \mathbf{Z}_{iv}^T \boldsymbol{\gamma}_2 + \varepsilon_{iv}$, $X_{ip} = \gamma_0 + \mathbf{W}_{ip}^T \boldsymbol{\gamma}_1 + \mathbf{Z}_{ip}^T \boldsymbol{\gamma}_2 + \varepsilon_{ip}$, $E(\varepsilon_{iv}) = E(\varepsilon_{ip}) = 0$ and $\text{var}(\varepsilon_{iv}) = \text{var}(\varepsilon_{ip}) = \sigma_\varepsilon^2$. As it turns out, all the methods to be discussed in this paper are numerically unchanged if we add the assumption that ε_{iv} and ε_{ip} are also normally distributed. We write this as follows.

Assumption 1 The mean of X given $(\mathbf{W}, \mathbf{Z})$ equals $\gamma_0 + \mathbf{W}^T \boldsymbol{\gamma}_1 + \mathbf{Z}^T \boldsymbol{\gamma}_2$ and the variance of X given $(\mathbf{W}, \mathbf{Z})$ equals σ_ε^2 in both the primary and the validation data.

The second part of the model is the relationship between Y and $(X, \mathbf{W}, \mathbf{Z})$ in the primary data. Here it is typical to make two assumptions.

Assumption 2 The measurement error is non-differential, i.e., $\mathbf{W}$ is independent of Y given $(X, \mathbf{Z})$.

Assumption 3 In the primary data, the distribution of Y given $(X, \mathbf{Z})$ follows an additive model, with density or mass function $f(Y, \beta_0 + X \beta_1 + \mathbf{Z}^T \boldsymbol{\beta}_2, \boldsymbol{\zeta})$, where $\boldsymbol{\zeta}$ is a vector of nuisance parameters.

1.3. Regression calibration approximation and approaches to comparisons

Our goal is to compare the two-stage method and the regression calibration method. There are three ways that this can be done.

- First, one can do simulation studies, something we take up in Section 3.2.
- The second approach is theoretical, involving asymptotic theory, but it involves an approximation as used by Weller *et al.* [1] among many other authors. The approximation, called the *regression calibration* approximation, is stated below as Assumption 4. It allows for a clear asymptotic comparison among the methods, see Appendix A.
- The third approach is more formal than the other two and does not depend on the regression calibration approximation. Based only on Assumptions 1–3, one can develop the asymptotic distribution of the two-stage and regression calibration methods. We show how to do this for logistic regression in Appendix B. There are however problems with this in terms of numerical comparisons.
 - The two methods generally estimate different quantities, see Section 1.4 for cases that they do estimate the same thing.
 - Both estimators converge to quantities that are solutions to nonlinear integral equations which depend on the joint distribution of $(X, \mathbf{W}, \mathbf{Z})$. This makes numerical comparison of their biases and variances difficult because in each case, an integral equation must be solved.
 - Interestingly, we show in Appendix B that if we make the regression calibration approximation, the results of the developed correct asymptotic theory reduce to those in Appendix A.

The regression calibration approximation, which we state here as Assumption 4, has to do with the relationship between Y and $(\mathbf{W}, \mathbf{Z})$, i.e., the distribution of the *observed* data.

Assumption 4: The distribution of Y given $(\mathbf{W}, \mathbf{Z})$ is the same as that of Y given $(X, \mathbf{Z})$, except X is replaced by its regression on $(\mathbf{W}, \mathbf{Z})$, $E(X|\mathbf{W}, \mathbf{Z})$, and the intercept is changed [3]. In symbols, it is assumed that the density/mass function of Y given $(\mathbf{W}, \mathbf{Z})$ is $f(Y, \alpha_0 + \beta_1 E(X|\mathbf{W}, \mathbf{Z}) + \mathbf{Z}^T \boldsymbol{\beta}_2, \boldsymbol{\zeta}) = f(Y, \alpha_0 + \mathbf{W}^T \boldsymbol{\alpha}_1 + \mathbf{Z}^T \boldsymbol{\alpha}_2, \boldsymbol{\zeta})$, where $\boldsymbol{\alpha}_1 = \beta_1 \boldsymbol{\gamma}_1$ and $\boldsymbol{\alpha}_2 = \boldsymbol{\beta}_2 + \beta_1 \boldsymbol{\gamma}_2$.

1.4. Discussion of assumptions

It is important to emphasize that, for most nonlinear regressions including logistic regression, Assumptions 1–4 are contradictory. In particular, Assumptions 1–3 are the ones usually made in the analysis, from which not only does Assumption 4 not follow, but, strictly speaking, it is incompatible with them. There are at least three notable exceptions.

- If $\beta_1 = 0$, then the regression calibration approximation holds exactly and Assumption 4 is true. In this case, both the two-stage method and standard regression calibration method estimate $\beta_1 = 0$ asymptotically. It is a consequence of the work in Appendix A that the two estimates are asymptotically equivalent when $\beta_1 = 0$.
- The second case occurs when Y given $(X, \mathbf{Z})$ follows a homoscedastic linear model, in which case Assumption 4 is strictly true. As it turns out, in this case the two-stage method and the regression calibration method are asymptotically equivalent, as we show in Appendix A.
- The third is probit regression, where the probability of disease is $\text{pr}(Y=1|X, \mathbf{Z}) = \Phi(\beta_0 + X\beta_1 + \mathbf{Z}^T \boldsymbol{\beta}_2)$. If X given $(\mathbf{W}, \mathbf{Z})$ is normally distributed, then the observed data follow the probit model

$$\text{pr}(Y=1|\mathbf{W}, \mathbf{Z}) = \Phi(\beta_0^* + \mathbf{W}^T \boldsymbol{\gamma}_1 \beta_1^* + \mathbf{Z}^T \boldsymbol{\beta}_2^*),$$

where $\beta_1^* = \beta_1 / (1 + \beta_1^2 \sigma_\varepsilon^2)^{1/2}$, $\beta_0^* = (\beta_0 + \gamma_0 \beta_1) / (1 + \beta_1^2 \sigma_\varepsilon^2)^{1/2}$ and

$\boldsymbol{\beta}_2^* = (\boldsymbol{\beta}_2 + \beta_1 \boldsymbol{\gamma}_2) / (1 + \beta_1^2 \sigma_\varepsilon^2)^{1/2}$. In this probit case, maximum likelihood for the observed data of Y given $(\mathbf{W}, \mathbf{Z})$, the two-stage method and the standard regression calibration method all estimate β_1^* , and our asymptotic results apply. Of course, if $\beta_1^2 \sigma_\varepsilon^2$ is large, then all methods are therefore badly biased.

Outside of these important cases, we are left with a dilemma about the relevance of the asymptotic theory based on Assumption 4 for comparing the two methods. There are three possible ways to argue the relevance of the asymptotic theory based on Assumption 4, all of which are equivalent for our calculations but are very different conceptually. The first is the conventional approach of staying with the usually made Assumptions 1–3 and using

Assumption 4 as an approximation. In cases where $\beta_1^2 \sigma_\varepsilon^2$ is relatively small, this approximation is quite good and the theoretical results are relevant. One should be careful though and not extend this theory to the full parametric space. In the absence of relevant theoretical results when Assumption 4 provides a poor approximation, the statistical properties of the estimated slope β_1 should be investigated by extensive finite-sample simulations.

The second possibility is to make Assumptions 1, 2 and 4. In this case, the true model remains unknown, but one assumes that β_1 has some approximate relevance to the parameter

of interest, although we do not know exactly how to specify this relationship. The third possibility is to make Assumptions 1, 3 and 4. In this case, one does not assume that error is non-differential, but rather that it is such that Assumptions 1, 3 and 4 hold. Although all three possibilities are interesting, and our work applies to all of them, we will concentrate on the first one which is usually used in practical applications.

1.5. Summary and outline

In summary, a generalization of the framework of Weller *et al.* based on Assumptions 1–3 and using Assumption 4 as an approximation is as follows.

$$X_{iv} = \gamma_0 + \mathbf{W}_{iv}^T \boldsymbol{\gamma}_1 + \mathbf{Z}_{iv}^T \boldsymbol{\gamma}_2 + \varepsilon_{iv}; \quad (1)$$

$$\varepsilon_{iv} = \text{Normal}(0, \sigma_\varepsilon^2); \quad (2)$$

$$\left[Y_{ip} | \mathbf{W}_{ip}, \mathbf{Z}_{ip} \right] = f(Y_{ip}, \alpha_0 + \beta_1 \mathbf{W}_{ip}^T \boldsymbol{\gamma}_1 + \mathbf{Z}_{ip}^T \boldsymbol{\alpha}_2, \zeta). \quad (3)$$

The goal is to estimate β_1 . Model (1)–(3) represents a likelihood formulation for the observed data. This is a new insight, and it suggests immediately that the data can be fit, efficiently, by maximum likelihood to get an asymptotically efficient estimate of the crucial parameter β_1 .

An outline of this paper is as follows. In Section 2, we describe the two-stage method of Weller *et al.* generalized to our context (1)–(3). We then state our three main results: (a) in the linear model for Y given $(X, \mathbf{Z})$, all three methods are asymptotically equivalent; (b) maximum likelihood for (1)–(3) and the two-stage method are asymptotically equivalent; and (c) standard regression calibration is not in general asymptotically equivalent to maximum likelihood for (1)–(3).

In Section 3.1, we take up the issue of how much more efficient the two-stage method is theoretically compared to standard regression calibration, by studying asymptotic variances over a broad range of interesting problems, and via simulations. In contrast to Weller *et al.*, who use a single data set and compute estimated asymptotic variances, we find that in practice, the difference in asymptotic efficiency between the two methods typically appears to be slight, although there are some exceptions. As mentioned above, under usually made Assumptions 1–3, model (1)–(3) is merely an approximation; this suggests that a finite-sample simulation study that compares the standard regression calibration and two-stage methods be undertaken. The results of the simulation study are reported in Section 3.2. and Appendix C. We conclude with a discussion in Section 4. A sketch of the technical details when Assumption 4 approximately holds is given in Appendix A, while in Appendix B we derive the exact asymptotic distributions for logistic regression, and show that they reduce to the results in Appendix A when the regression calibration approximation is invoked.

2. Methods and results

2.1. Methods

Maximum likelihood for model (1)–(3) is straightforward. Standard regression calibration consists of first estimating $(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$ using the validation data only, substituting them into (3), assuming that $\boldsymbol{\gamma}_1$ is fixed to this estimated value and known, and then performing maximum likelihood to estimate the unknown slope β_1 .

The two-stage method is somewhat more complex. Again, Weller *et al.* first estimate (γ_1, γ_2) using the validation data to obtain $(\hat{\gamma}_1, \hat{\gamma}_2)$. Then estimation of β_1 proceeds in two stages. At stage 1, the following model is fit to the data:

$$[Y_{ip} | \mathbf{W}_{ip}, \mathbf{Z}_{ip}] = f(Y_{ip}, \alpha_0 + \mathbf{W}_{ip}^T \boldsymbol{\alpha}_1 + \mathbf{Z}_{ip}^T \boldsymbol{\alpha}_2, \zeta), \quad (4)$$

leading to estimates $\hat{\mathbf{a}}_1$ of the “naive” slopes. Because of model (3), one has $\beta_1 \mathbf{e}_{\text{one}} = \mathbf{a}_1 // \boldsymbol{\gamma}_1$, where $//$ denotes element-wise division and where $\mathbf{e}_{\text{one}} = (1, 1, \dots, 1)^T$. At stage 2, Weller *et al.* estimate β_1 by forming the weighted linear combination $\boldsymbol{\tau}^T(\mathbf{a}_1 // \boldsymbol{\gamma}_1)$ and choosing the weights to minimize the asymptotic variance of this linear combination subject to the constraint that $\mathbf{e}_{\text{one}}^T \boldsymbol{\tau} = 1$.

2.2. Non-invariance of the two-stage method

Let $\mathbf{\Lambda}$ be any full rank square matrix whose number of columns is the dimension of $\mathbf{W}$.

Consider using the transformed vector of surrogates $\mathbf{W}_{iv}^* = \mathbf{\Lambda} \mathbf{W}_{iv}$ and $\mathbf{W}_{ip}^* = \mathbf{\Lambda} \mathbf{W}_{ip}$ based on the same full-rank linear transformation $\mathbf{\Lambda}$ applied to both the validation and primary data. It is easy to see theoretically that both maximum likelihood and regression calibration give exactly the same estimate of β_1 whether $(\mathbf{W}_{iv}, \mathbf{W}_{ip})$ or $(\mathbf{W}_{iv}^*, \mathbf{W}_{ip}^*)$ are used. For example, with regression calibration, dropping precisely measured covariates for simplicity, if $\hat{\boldsymbol{\gamma}}_1$ is the estimate of $\boldsymbol{\gamma}_1$ based on the linear regression of X_{iv} on $\mathbf{W}_{iv}$, then $\hat{\boldsymbol{\gamma}}_1^* = (\mathbf{\Lambda}^T)^{-1} \hat{\boldsymbol{\gamma}}_1$ is the corresponding estimate if $\mathbf{W}_{iv}^*$ is used, and the regression calibration predictions are equivalent, i.e., $\mathbf{W}_{ip}^T \hat{\boldsymbol{\gamma}}_1 = \mathbf{W}_{ip}^{*T} \hat{\boldsymbol{\gamma}}_1^*$. This is as it should be: a 1–1 transformation of nuisance parameters should not affect the estimate of the main parameter of interest.

In contrast, the two-stage estimator enjoys no such property, and indeed in some settings it can be made to take on almost any value by varying $\mathbf{\Lambda}$. For example, at

<http://www.stat.tamu.edu/~carroll/TestDataSetprimary.txt>

<http://www.stat.tamu.edu/~carroll/TestDataSetvalidation.txt>

can be found primary and validation data sets when $\mathbf{W}$ is two dimensional. We generated 10,000 full rank linear transformations $\mathbf{\Lambda}$ by having their entries be standard normal random variables. The two-stage estimator for the transformed $\mathbf{W}$ -variables over these 10,000 simulated full rank linear transformations had a minimum = −36.11, a maximum = 2.43, a mean = 0.79 and a 98% confidence interval of [0.75, 0.83].

We suggest linearly transforming $\mathbf{W}$ in the validation study to have mean zero and identity covariance matrix, applying the same transformation to the primary data, and then applying the two-stage method. In our extensive simulations, we have found that the modified two-stage method provides better results than the original two-stage estimator.

2.3. Assumptions and results

We show that the two-stage method achieves optimal asymptotic properties under the regression calibration assumption.

Here is our main result, the proof of which is given in Appendix A. The asymptotic variances for maximum likelihood, regression calibration and the two-stage method are given in Appendix A at equations (A.3), (A.4) and (A.5), respectively.

Theorem 1—Under model (1)–(3), the maximum likelihood and standard regression calibration estimators have the same asymptotic variances if Y given (X, Z) follows a

homoscedastic linear model or if $\beta_1 = 0$, but not in general. Under model (1)-(3), the two-stage estimator of β_1 is asymptotically equivalent to maximum likelihood if there are no nuisance parameters ζ or if Y given $(\mathbf{W}, \mathbf{Z})$ follows a generalized linear model in the primary data.

2. Methods Comparisons

3.1. Asymptotic efficiency comparisons under the regression calibration assumption

3.2. Background

While, at least technically, regression calibration is not efficient for approximate model (1)-(3), it remains to be seen how inefficient it is in realistic contexts. The asymptotic variances for all considered estimators are given at (A.3), (A.4) and (A.5) in Appendix A.

To compare these variances, we consider the case of logistic regression. Without loss of generality one can set $E(X) = 0$ and $\text{var}(X) = \boldsymbol{\gamma}^T \text{cov}(\mathbf{W}) \boldsymbol{\gamma} + \sigma_\varepsilon^2 = 1$, and because maximum likelihood, regression calibration and the modified two-stage method for estimating β_1 are all invariant to full-rank linear transformations of $\mathbf{W}$, without loss of generality one can assume that $\text{cov}(\mathbf{W}) = \mathbf{I}$. Assuming that $\varepsilon_{ip} = \varepsilon_{ip} = \text{Normal}(0, \sigma_\varepsilon^2)$, the only items that one needs to vary are (a) the distribution of $\mathbf{W}$; (b) $(\sigma_\varepsilon^2, \boldsymbol{\gamma})$ subject to $\boldsymbol{\gamma}^T \boldsymbol{\gamma} + \sigma_\varepsilon^2 = 1$; (c) the odds ratio of interest; and (d) the rate of disease in the source population. The latter two determine the underlying logistic intercept β_0 and the slope β_1 .

3.1.2. Normal and Chi-square $\mathbf{W}$ —First, $\mathbf{W}$ was set to be bivariate, with mean zero and covariance matrix the identity, but generated from (a) a bivariate normal distribution and then rescaled; (b) a bivariate chi-square distribution with one degree of freedom and then rescaled; and (c) a bivariate chi-square distribution with one degree of freedom with a random sign, resulting in a mixture distribution. Without loss of generality, we defined $X = \mathbf{W}^T \boldsymbol{\gamma} + \varepsilon$ to have mean zero and variance one. We did this by first generating $\sigma_\varepsilon^2 = \text{Uniform}(0, 1)$. We then generated $\boldsymbol{\gamma}^* = \text{Normal}(0, \mathbf{I}_2)$, and defined $\boldsymbol{\gamma} = (1 - \sigma_\varepsilon^2)^{1/2} \boldsymbol{\gamma}^* / (\boldsymbol{\gamma}^{*T} \boldsymbol{\gamma}^*)^{1/2}$. Then we set the true logistic intercept β_0 and the true logistic slope β_1 so that the odds ratio of moving from the 10th to the 90th percentile of the distribution of X was in the set (2.0, 3.0, 4.0, 5.0, 6.0), and the probability of disease in the source population was in the set (0.075, 0.150, ..., 0.450). Hence, we considered relatively rare to common diseases, and relatively modest to large true odds ratios, along with fairly extreme distributions for $\mathbf{W}$. We varied the ratio of the validation sample size to the primary sample size, $c_F = n_v/n_p$, to be (0.01, 0.10). For each of these settings, we searched over 1,000 cases of $(\sigma_\varepsilon^2, \boldsymbol{\gamma})$. When computing asymptotic variances, we set $\alpha_0 = \beta_0 + \beta_1^2 \sigma_\varepsilon^2 / 2$, the usual regression calibration approximation. Monte-Carlo was used to calculate integrals where required.

The results are striking: in all cases, the asymptotic variance of the standard regression calibration estimator of β_1 was less than 1.5% larger than the asymptotic variance of the maximum likelihood estimator, and the two-stage estimator.

3.1.3. Lognormal and Gamma $\mathbf{W}$ —We also considered lognormal cases. We let Σ_v be a correlation matrix with correlation ρ in the off-diagonal element. We generated 1,000 settings where we generated $-0.95 \leq \rho \leq 0.95$, σ_ε^2 and $\boldsymbol{\gamma}$ so that $\text{var}(X) = 1$, see Section 3.1.2. We set $\alpha_0 = \beta_0 + \beta_1^2 \sigma_\varepsilon^2 / 2$, where β_0 and β_1 are defined below so that the rate of disease and the odds ratio are the same as in Section 3.1.2.

The $\mathbf{W}$ data were generated as follows. We describe here the theoretical setting, although simulation was actually used in all cases to compute the integrals involved in the asymptotic variances. First, $\mathbf{W}_*$ were bivariate independent lognormal random variables such that $\log(\mathbf{W}_*)$ are bivariate independent normal variables with mean zero and variance $\zeta = \text{Uniform}(0,1)$. We defined $\mathbf{W}_{**}$ to be the standardized version of $\mathbf{W}_*$, with both independent variables having mean zero and variance one, then made $\mathbf{W}$ to be a linear transformation of $\mathbf{W}_{**}$ that had covariance matrix Σ_v and exactly mean zero. In symbols, let $\mathbf{Z}$ be a bivariate standard normal random variable. Then

$$\begin{aligned} \mathbf{W}_{i*} &= \exp(\zeta^{1/2} \mathbf{Z}_i), \quad i=1, 2; \\ \mathbf{W}_{i**} &= \{\mathbf{W}_{i*} - \exp(\zeta/2)\} / \{\exp(2\zeta) - \exp(\zeta)\}^{1/2}, \quad i=1, 2; \\ \mathbf{W} &= \sum_v^{1/2} \{\mathbf{W}_{**} - E(\mathbf{W}_{**})\}. \end{aligned}$$

The lognormal distribution is notorious for generating a few extremely wild observations. For example, in Figures 1 and 2 we give box plots of 5 randomly generated data sets of size 10,000 from the lognormal and gamma distributions, each with the coefficient of variation equal to 1.0. These are plotted on the same scale to make the point that the lognormal distribution is much more prone to wild observations than is the gamma distribution. Thus, the distribution that we actually used for $\mathbf{W}$ was a truncated one, truncated at $4 \times \text{IQR}$ above the 75th or below the 25th percentile, where IQR is the interquartile range. Using this rule of thumb, in the lognormal case we removed 1.18% of the distribution of $\mathbf{W}$, while for the gamma distribution the corresponding figure is 0.24%. We then generated X from the remaining observations. We set β_1 so that the odds ratio for moving from the 10th to the 90th percentile of the distribution of X would be a predefined value OR , and then formed β_0 so that the probability $\text{pr}(Y=1)$ of disease in the population would be equal to a predefined value π .

In most interesting situations for practical applications, with odds ratios $OR = 6$, the worst case for regression calibration was when $\pi = 0.075$, $OR = 6$, and $n_v/n_p = 0.01$. In the 1,000 cases, the maximum asymptotic variance inflation of standard regression calibration was 13%, and more than 90% of these test cases had less than a 2% variance inflation. If $n_v/n_p = 0.10$, there was at worst a 1% variance inflation. In other words, there was little or no asymptotic variance inflation in these cases. We have done calculations without truncating the lognormal distribution, and the variance inflation becomes larger, sometimes much larger, reaching 33% in extreme cases, but these situations seem to be unrealistic in practice because of the wild observations. In the gamma case, there was essentially no variance inflation.

We have also checked the lognormal case should the odds ratio for moving from the 10th to the 90th percentile of the distribution of X reach a (practically impossible in our experience) value of $OR = 45$. The maximum variance inflation of standard regression calibration was 83% with truncation and 147% without. As was mentioned previously, in all those cases, the regression calibration approximation leads to a severely misspecified model. The resulting slopes estimated by both methods involve large biases and should not be used for measurement error adjustment.

3.2. Finite-sample simulations

All results up to now have been asymptotic results based on the regression calibration approximation in equations (1)–(3). In this section, we use simulations to examine the finite sample properties of the two-stage and regression calibration methods when the regression calibration assumption given by equation (3) is violated. We especially concentrate on cases

when equation (3) does not hold even approximately, and the relevance of asymptotic variance calculations under this assumption becomes questionable.

3.2.1. Normal $\mathbf{W}$ —We let Y be a binary disease outcome, X = true exposure, W_1 = first surrogate measurement of X and W_2 = second surrogate measurement of X . We performed 5 sets of 10,000 simulations. For each simulation, we generated two independent data sets, a main study and a validation study. In the main study, each subject had variables Y , W_1 and W_2 . In the validation study, each subject had the variables X , W_1 and W_2 . The sample size for the main study was 10,000, while the sample size for the calibration study was 1,000. We generated the data according to the following models. First, (W_1, W_2) was bivariate normal with zero means, common variance equal to 1.0, and correlation ρ . We generated $\varepsilon \sim \text{Normal}(0, \sigma_\varepsilon^2)$ and set $X = \gamma_0 + \gamma_1 W_1 + \gamma_2 W_2 + \varepsilon$. Finally, we generated Y as a binary variable with $\text{pr}(Y=1/X) = H(\beta_0 + \beta_1 X)$, where $H(\cdot)$ is the logistic distribution function. In all simulations, we set $\rho = 0.5$, $\gamma_0 = 0$, and $\sigma_\varepsilon^2 = 1 - (\gamma_1^2 + \gamma_2^2 + 2\rho\gamma_1\gamma_2)$, thus ensuring that $\text{var}(X) = 1$. Then we set the true logistic intercept β_0 and the true logistic slope β_1 so that the odds ratio for increasing X by two standard deviations of its distribution was equal to a predefined value OR, and the probability of disease in the source population was equal to a predefined value π . Table 1 shows the values of the parameters for each simulation. For each simulation, we estimated β_1 using standard regression calibration and both the original and modified version of the two-stage method of Weller *et al.* Table 2 shows the results of the simulations.

The results are instructive. In cases 1 – 4, the empirical variance of standard regression calibration is equal to or smaller than that of the unmodified two-stage method with essentially no difference between standard regression calibration and the modified two-stage method. In the most extreme case 5, there is a 3% and 9% difference in variance, with the original and modified two-stage methods, respectively, being less variable. However, in all cases, standard regression calibration is less biased than either two-stage method and, as a result, has smaller root mean squared error. When one deals with true odds ratios greater than 20, cases 4–5, the regression calibration approximation can be particularly poor, and finite-sample properties of the considered estimators become especially relevant.

3.2.2. Lognormal $\mathbf{W}$ —We also ran simulations when surrogate measurements had a lognormal distribution. In addition to a normally distributed error ε in the linear regression of X on $\mathbf{W}$, at the request of one of the reviewers, we also considered the case of a lognormal error, both homoscedastic and heteroscedastic that depended on $\mathbf{W}$. The description of those simulations and their results are given in Appendix C. All results are reasonably in line with those in the previous section for the normal case.

3.2.3. Categorical $\mathbf{W}$ —We used the data described by Weller *et al.* to understand the properties of the methods when W_1, W_2 are categorical. Following Section 6 in Weller *et al.*, we considered the case when $\mathbf{W} = (W_1, W_2)$ had in total 3 distinct values. We set $\mathbf{W}$ to be categorical with 3 levels according to probabilities (0.4, 0.3, 0.3), and let W_1, W_2 be the dummy variable indicators according to the two highest levels, with the induced correlation of $\rho = -3/7$. We set $\gamma_0 = 0.127$, and $\sigma_\varepsilon^2 = 1 - (\gamma_1^2 + \gamma_2^2 + 2\rho\gamma_1\gamma_2)$. Table 3 shows the values of the parameters for each simulation. In the simulations, the primary sample size was 10,000, the validation sample size was 1,000, and 10,000 simulated data sets were generated at each configuration. The results given in Table 4 are quantitatively similar to those for continuously distributed surrogate values.

4. Discussion

We have shown that the two-stage method of Weller *et al.* is not regular in the sense that the finite sample estimates of the odds ratio can vary widely if different 1–1 transformations of nuisance parameters are made. Despite this, the two-stage method is asymptotically equivalent to maximum likelihood for approximate model (1)–(3).

Our most striking conclusion is that, under the regression calibration approximation, the two-stage method is asymptotically less or no more variable than standard regression calibration in a wide variety of circumstances.

We performed numerical calculations that somewhat temper the main conclusion. We found in most cases that the two-stage and standard regression calibration methods were very nearly equivalent in asymptotic variance, exceptions being when the odds ratio was very and perhaps unrealistically large. It would be interesting to find practical cases with modest odds ratios and well behaved distributions of $\mathbf{W}$ when the variance gain from using the two-stage method was substantial, but we were unable to find such cases.

We have also conducted an extensive simulation study to compare finite-sample properties of the estimators, with a special attention to cases where the regression calibration assumption does not provide a good approximation and the relevance of the asymptotic variances of the estimators under this approximation becomes questionable. We have found that only in such extreme cases the standard regression calibration estimator had a slightly greater finite-sample variance than the two-stage estimator. Yet, in all cases, standard regression calibration was less biased and, as a result, had smaller root mean squared error.

Under the regression calibration approximation, the two-stage method is asymptotically efficient for the problems we have described. It would be interesting to see how the method could be applied to other problems and assess its asymptotic properties in these situations. One of the properties of standard regression calibration, or maximum likelihood in the regression calibration context, is that it generalizes easily to cases where there are more than one predictor measured with error, and the surrogates for each such predictor may or may not be the same. It is possible to define a two-stage procedure for such cases, although it is not straightforward and the resulting estimators remain not invariant to reparameterization of nuisance parameters. Whether they have optimality properties that match the asymptotic properties of the regression calibration maximum likelihood is unknown.

To illustrate this, suppose that there are $K = 2$ variables measured with error and $X_1 = \mathbf{W}_1^T \boldsymbol{\gamma}_1 + \mathbf{W}_2^T \boldsymbol{\gamma}_2 + \varepsilon_1$ and $X_2 = \mathbf{W}_2^T \boldsymbol{\gamma}_3 + \mathbf{W}_3^T \boldsymbol{\gamma}_4 + \varepsilon_2$. This is a case that one set of covariates, $\mathbf{W}_2$ appear in each regression, but the others, $\mathbf{W}_1$ and $\mathbf{W}_3$, appear in only one of the regressions. Then write $\mathbf{W}^T = (\mathbf{W}_1^T, \mathbf{W}_2^T, \mathbf{W}_3^T)$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)^T$, so that

$$E(X_1 | \mathbf{W}_1, \mathbf{W}_2) \boldsymbol{\beta}_1 + E(X_2 | \mathbf{W}_2, \mathbf{W}_3) \boldsymbol{\beta}_2 = \begin{bmatrix} \mathbf{W}_1^T & \mathbf{W}_2^T & \mathbf{W}_3^T \\ \mathbf{W}_1^T & \mathbf{W}_2^T & \mathbf{W}_3^T \end{bmatrix} \begin{bmatrix} \boldsymbol{\gamma}_1 & \mathbf{0} \\ \boldsymbol{\gamma}_2 & \boldsymbol{\gamma}_3 \\ \mathbf{0} & \boldsymbol{\gamma}_4 \end{bmatrix} \boldsymbol{\beta}.$$

It is not clear what $\boldsymbol{\alpha}$ is here, and how one can take advantage of the fact that there is knowledge that $\mathbf{W}_3$ is not a predictor of X_1 given $(\mathbf{W}_1, \mathbf{W}_2)$. While it is possible to devise estimators, and weight them, their properties are unclear; maximum likelihood and standard regression calibration have no such lack of clarity.

Acknowledgments

We thank Dr. Weller for providing us with the data used in their original analysis. Carroll's research was supported by a grant from the National Cancer Institute (CA57030) and by Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST).

References

1. Weller EA, Milton DK, Eisen EA, Spiegelman D. Regression calibration for logistic regression with multiple surrogates for one exposure. *Journal of Statistical Planning and Inference*. 2007; 137:449–461.
2. Carroll RJ, Stefanski LA. Approximate quasi-likelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association*. 1990; 85:652–663.
3. Carroll, R.J.; Ruppert, D.; Stefanski, L.A.; Crainiceanu, CM. *Measurement Error in Nonlinear Models: A Modern Perspective*. 2. Chapman and Hall CRC Press; Boca Raton, FL: 2006.
4. Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risks and confidence intervals for systematic within-person measurement error. *Statistics in Medicine*. 1989; 8:1051–1070. [PubMed: 2799131]
5. Thurston SW, Spiegelman D, Ruppert D. Equivalence of regression calibration methods in main study/external validation study designs. *Journal of Statistical Planning and Inference* 2003. 2003; 113:527–539.

Appendix A: Sketch of Technical Arguments Under the Regression Calibration Model

A.1. Simplifications

It is readily seen that the scores for estimating $(\alpha, \beta_1, \beta_2, \gamma_0, \gamma_1, \gamma_2, \zeta)$ in model (3) are orthogonal to the score for estimating σ_e , so we will without loss of generality assume σ_e^2 is known. Also, without loss of generality, we center X_{iv} so that $E(X_{iv}) = 0$. We also center $\mathbf{W}_{iv}$ by subtracting a constant vector so that $E(\mathbf{W}_{iv}) = 0$, and subtract the same constant from $\mathbf{W}_{ip}$; the same procedure is applied to $\mathbf{Z}_{iv}$ and $\mathbf{Z}_{ip}$. This means that $\gamma_0 = 0$ and we can ignore it in our development.

It suffices to prove the main result in the special case that there are no nuisance parameters ζ . In generalized linear models there may be a nuisance parameter ζ , but the loglikelihood is of the form

$$\frac{a(Y)b(\alpha+\beta_1\mathbf{W}^T\boldsymbol{\gamma}_1+\mathbf{Z}^T\boldsymbol{\beta}_2)-c(\alpha+\beta_1\mathbf{W}^T\boldsymbol{\gamma}_1+\mathbf{Z}^T\boldsymbol{\beta}_2)}{\zeta}+d(Y,\zeta).$$

However, the score for ζ is orthogonal to the score for $(\alpha, \beta_1, \gamma_1, \beta_2)$. To see this well-known fact, we note that the score for $(\alpha, \beta_1, \gamma_1, \beta_2)$ is

$$\mathbf{G}(Y, \mathbf{W}, \mathbf{Z}) = (1, \mathbf{W}^T\boldsymbol{\gamma}_1, \beta_1\mathbf{W}^T, \mathbf{Z}^T)^T \times \frac{a(Y)b^{(1)}(\alpha+\beta_1\mathbf{W}^T\boldsymbol{\gamma}_1+\mathbf{Z}^T\boldsymbol{\beta}_2)+c^{(1)}(\alpha+\beta_1\mathbf{W}^T\boldsymbol{\gamma}_1+\mathbf{Z}^T\boldsymbol{\beta}_2)}{\zeta},$$

where the superscript means first derivative. However, because this is a generalized linear model, $E\{\mathbf{G}(Y, \mathbf{W}, \mathbf{Z})/\mathbf{W}, \mathbf{Z}\} = 0$. Differentiating the score for $(\alpha, \beta_1, \gamma_1, \beta_2)$ with respect to ζ yields $-\mathbf{G}(Y, \mathbf{W}, \mathbf{Z})/\zeta^2$, so it too has mean zero. This means that the Fisher information

matrix is block-diagonal, that the scores are orthogonal, and hence in generalized linear models the asymptotic theory is the same as if δ were known.

It suffices to work in the special case with no accurately measured covariates $\mathbf{Z}$ in the model: the extra generality is accomplished by somewhat more complicated definitions of what are terms $\mathbf{A}$ and $\mathbf{B}$ defined below. In this case, we replace $\boldsymbol{\gamma}_1$ by $\boldsymbol{\gamma}$ and β_1 by β . The log-likelihood score then takes the form $\ell(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \log\{f(\boldsymbol{\gamma}, \boldsymbol{\beta})\} / \sqrt{v}$ with $v = \alpha + \boldsymbol{\beta}^T \mathbf{W} \boldsymbol{\gamma}$. Hence, we do the detailed calculations for the model

$$\begin{aligned} X_{iv} &= \mathbf{W}_{iv}^T \boldsymbol{\gamma} + \varepsilon_{iv}; \\ \varepsilon_{iv} &= \text{Normal}(0, \sigma_\varepsilon^2); \\ \text{cov}(\mathbf{W}_{iv}) &= \Sigma_v; \\ [Y_{ip} | \mathbf{W}_{ip}] &= f(Y_{ip}, \alpha + \boldsymbol{\beta}^T \mathbf{W}_{ip} \boldsymbol{\gamma}). \end{aligned}$$

We also use standard facts about likelihood theory, e.g., that the square of the score has expectation that is -1 times the expectation of the derivative of the score, so that, for example, we define in the primary study

$$\begin{aligned} \mathbf{A} &= E\{\ell^2(\cdot)\} = -E\{\ell^{(1)}(\cdot)\}; \\ \mathbf{B} &= E\{\mathbf{W} \ell^2(\cdot)\} = -E\{\mathbf{W} \ell^{(1)}(\cdot)\}; \\ \mathbf{C} &= E\{\mathbf{W} \mathbf{W}^T \ell^2(\cdot)\} = -E\{\mathbf{W} \mathbf{W}^T \ell^{(1)}(\cdot)\}. \end{aligned}$$

These are easily confirmed by conditioning on $\mathbf{W}$. Define $n_v/n_p = c_F$

A.2. Some Matrix Formulae

We use repeatedly the fact that if

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_{11} & \mathbf{a}_{12} \\ \mathbf{a}_{21} & \mathbf{a}_{22} \end{bmatrix},$$

then

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{b}_{11} & \mathbf{b}_{12} \\ \mathbf{b}_{21} & \mathbf{b}_{22} \end{bmatrix},$$

where $\mathbf{b}_{22} = (\mathbf{a}_{22} - \mathbf{a}_{21} \mathbf{a}_{11}^{-1} \mathbf{a}_{12})^{-1}$, $\mathbf{b}_{11} = \mathbf{a}_{11}^{-1} + \mathbf{a}_{11}^{-1} \mathbf{a}_{12} \mathbf{b}_{22} \mathbf{a}_{21} \mathbf{a}_{11}^{-1}$, $\mathbf{b}_{12} = -\mathbf{a}_{11}^{-1} \mathbf{a}_{12} \mathbf{b}_{22}$, and $\mathbf{b}_{21} = -\mathbf{b}_{22} \mathbf{a}_{21} \mathbf{a}_{11}^{-1}$. It is also worth remembering that

$$(\mathbf{a} + \mathbf{b}^{-1})^{-1} = \mathbf{a}^{-1} - \mathbf{a}^{-1} (\mathbf{a}^{-1} + \mathbf{b})^{-1} \mathbf{a}^{-1}. \quad (\text{A.1})$$

A.3. Asymptotic Theory for Maximum Likelihood

The score for $(\alpha, \boldsymbol{\gamma}, \beta)$ is given as

$$\mathbf{S} = \begin{bmatrix} \sum_{i=1}^{n_p} \ell_i(\cdot) \\ \sum_{i=1}^{n_v} \mathbf{W}_{iv} (X_{iv} - \mathbf{W}_{iv}^T \boldsymbol{\gamma}) / \sigma_u^2 + \sum_{i=1}^{n_p} \beta \mathbf{W}_{ip} \ell_i(\cdot) \\ \sum_{i=1}^{n_p} \mathbf{W}_{ip}^T \boldsymbol{\gamma} \ell_i(\cdot) \end{bmatrix}$$

It follows with routine calculations, and remembering that $\text{cov}(\mathbf{W}_{iv}) = \boldsymbol{\Sigma}_v$, that

$$E(\mathbf{S}\mathbf{S}^T) = n_p \begin{bmatrix} A & \beta \mathbf{B}^T & \mathbf{B}^T \boldsymbol{\gamma} \\ \beta \mathbf{B} & (c_F / \sigma_u^2) \boldsymbol{\Sigma}_v + \beta^2 \mathbf{C} & \beta \mathbf{C} \boldsymbol{\gamma} \\ \mathbf{B}^T \boldsymbol{\gamma} & \beta \boldsymbol{\gamma}^T \mathbf{C} & \boldsymbol{\gamma}^T \mathbf{C} \boldsymbol{\gamma} \end{bmatrix} = n_p \tilde{\mathbf{J}}.$$

Let $\mathbf{e}_1^T = (0, 0, \dots, 0, 1)$. It follows from likelihood theory that

$$n_p^{1/2} (\hat{\boldsymbol{\beta}}_{\text{ML}} - \boldsymbol{\beta}) \rightarrow \text{Normal}(0, \sigma_{\text{ML}}^2 = \mathbf{e}_1^T \tilde{\mathbf{J}}^{-1} \mathbf{e}_1). \quad (\text{A.2})$$

We want to compute $\mathbf{e}_1^T \tilde{\mathbf{J}}^{-1} \mathbf{e}_1$. Define

$$\begin{aligned} \mathbf{a}_{11} &= \begin{bmatrix} A & \beta \mathbf{B}^T \\ \beta \mathbf{B} & (c_F / \sigma_u^2) \boldsymbol{\Sigma}_v + \beta^2 \mathbf{C} \end{bmatrix}; \\ \mathbf{a}_{21} &= \begin{bmatrix} \boldsymbol{\gamma}^T \mathbf{B} & \beta \boldsymbol{\gamma}^T \mathbf{C} \end{bmatrix}; \\ \mathbf{a}_{12} &= \mathbf{a}_{21}^T; \\ a_{22} &= \boldsymbol{\gamma}^T \mathbf{C} \boldsymbol{\gamma}. \end{aligned}$$

Then

$$\sigma_{\text{ML}}^2 = \mathbf{e}_1^T \tilde{\mathbf{J}}^{-1} \mathbf{e}_1 = (a_{22} - \mathbf{a}_{21} \mathbf{a}_{11}^{-1} \mathbf{a}_{12})^{-1} = (\boldsymbol{\gamma}^T \boldsymbol{\Lambda} \boldsymbol{\gamma})^{-1}; \quad (\text{A.3})$$

$$\boldsymbol{\Lambda} = \mathbf{C} - \begin{bmatrix} \mathbf{B} & \beta \mathbf{C} \end{bmatrix} \begin{bmatrix} A & \beta \mathbf{B}^T \\ \beta \mathbf{B} & (c_F / \sigma_u^2) \boldsymbol{\Sigma}_v + \beta^2 \mathbf{C} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{B}^T \\ \beta \mathbf{C} \end{bmatrix}. \quad (\text{A.4})$$

A.4. Asymptotic Theory for Regression Calibration

The estimating equation for $(\alpha, \boldsymbol{\gamma}, \beta)$ is

$$\begin{aligned}
0 &= \begin{bmatrix} n_p^{-1/2} \sum_{i=1}^{n_p} \ell(Y_{ip}, \widehat{\alpha} + \beta \mathbf{W}_{ip}^T \boldsymbol{\gamma}) \\ n_p^{-1/2} \sum_{i=1}^{n_p} \mathbf{W}_{iv} (X_{iv} - \mathbf{W}_{iv}^T \boldsymbol{\gamma}) \\ n_p^{-1/2} \sum_{i=1}^{n_p} \mathbf{W}_{ip}^T \boldsymbol{\gamma} \ell(Y_{ip}, \widehat{\alpha} + \beta \mathbf{W}_{ip}^T \boldsymbol{\gamma}) \end{bmatrix} \\
&= n_p^{-1/2} \begin{bmatrix} \sum_{i=1}^{n_p} \ell_i(\cdot) \\ \sum_{i=1}^{n_p} \mathbf{W}_{iv} (X_{iv} - \mathbf{W}_{iv}^T \boldsymbol{\gamma}) \\ \sum_{i=1}^{n_p} \mathbf{W}_{ip}^T \boldsymbol{\gamma} \ell_i(\cdot) \end{bmatrix} - \mathbf{D} n_p^{1/2} \begin{pmatrix} \widehat{\alpha} - \alpha \\ \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \\ \widehat{\beta} - \beta \end{pmatrix} + o_p(1),
\end{aligned}$$

where

$$\mathbf{D} = \begin{bmatrix} A & \beta \mathbf{B}^T & \mathbf{B}^T \boldsymbol{\gamma} \\ 0 & c_F \sum_v & 0 \\ \boldsymbol{\gamma}^T \mathbf{B} & \beta \boldsymbol{\gamma}^T \mathbf{C} & \boldsymbol{\gamma}^T \mathbf{C} \boldsymbol{\gamma} \end{bmatrix}.$$

Define $\mathbf{M}$ to be the covariance matrix of the left hand side of the last expansion so that

$$\mathbf{M} = \begin{bmatrix} A & 0 & \mathbf{B}^T \boldsymbol{\gamma} \\ 0 & \sigma_\varepsilon^2 c_F \sum_v & 0 \\ \boldsymbol{\gamma}^T \mathbf{B} & 0 & \boldsymbol{\gamma}^T \mathbf{C} \boldsymbol{\gamma} \end{bmatrix}.$$

Then we have that

$$n_p^{1/2} (\widehat{\beta}_{RC} - \beta) \rightarrow \text{Normal} \left\{ 0, \sigma_{RC}^2 = \mathbf{e}_1^T \mathbf{D}^{-1} \mathbf{M} (\mathbf{D}^{-1})^T \mathbf{e}_1 \right\}. \quad (\text{A.4})$$

A.5. Asymptotic Theory for the Two-Stage Method

This method estimates $\boldsymbol{\gamma}$ in the validation data only, and thus

$n_v^{1/2} (\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \rightarrow \text{Normal}(0, \sigma_\varepsilon^2 \sum_v^{-1})$. Hence, $n_p^{1/2} (\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \rightarrow \text{Normal}\{0, (\sigma_\varepsilon^2 / c_F) \sum_v^{-1}\}$. The method then fits the likelihood $f(Y_{ip}, \alpha_0 + \mathbf{W}_{ip}^T \boldsymbol{\alpha}_1)$, and the loglikelihood score for this is just $(1, \mathbf{W}_{ip})^T \ell(Y_{ip}, \alpha_0 + \mathbf{W}_{ip}^T \boldsymbol{\alpha}_1)$. This has information matrix

$$\mathfrak{I}_2 = \begin{bmatrix} A & \mathbf{B}^T \\ \mathbf{B} & \mathbf{C} \end{bmatrix}.$$

Hence, the asymptotic covariance matrix of $n_p^{1/2} (\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1)$ is the lower right corner submatrix of $\mathfrak{I}_2^{-1}$, and thus

$$n_p^{1/2} (\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1) \rightarrow \text{Normal} \left\{ 0, (\mathbf{C} - \mathbf{B} \mathbf{A}^{-1} \mathbf{B}^T)^{-1} \right\}.$$

The two-stage method considers a weighted linear combination of $\{\text{diag}(\hat{\gamma})\}^{-1}\hat{\alpha}_1$. Now, by algebra,

$$n_p^{1/2} \left[\{\text{diag}(\hat{\gamma})\}^{-1}\hat{\alpha}_1 - \{\text{diag}(\gamma)\}^{-1}\alpha_1 \right] = \{\text{diag}(\gamma)\}^{-1} n_p^{1/2} \{(\hat{\alpha}_1 - \alpha_1) - \beta(\hat{\gamma} - \gamma)\} + o_p(1),$$

remembering that $\beta(1, \dots, 1)^T = \{\text{diag}(\gamma)\}^{-1}\alpha_1$. Hence,

$$n_p^{1/2} \left[\{\text{diag}(\hat{\gamma})\}^{-1}\hat{\alpha}_1 - \{\text{diag}(\gamma)\}^{-1}\alpha_1 \right] \rightarrow \text{Normal}(0, \Xi);$$

$$\Xi = \{\text{diag}(\gamma)\}^{-1} \left\{ (\mathbf{C} - \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T)^{-1} + (\beta^2 \sigma_u^2 / c_F) \Sigma_v^{-1} \right\} \{\text{diag}(\gamma)\}^{-1}.$$

If $\mathbf{e}_{\text{one}} = (1, 1, \dots, 1)^T$, then the optimal weight is $\omega = \Xi^{-1} \mathbf{e}_{\text{one}} (\mathbf{e}_{\text{one}}^T \Xi^{-1} \mathbf{e}_{\text{one}})^{-1}$, from which it follows that $\omega^T \Xi \omega = (\mathbf{e}_{\text{one}}^T \Xi^{-1} \mathbf{e}_{\text{one}})^{-1}$, and thus

$$n_p^{1/2} (\hat{\beta}_{\text{ws}} - \beta) \rightarrow \text{Normal} \left\{ 0, (\mathbf{e}_{\text{one}}^T \Xi^{-1} \mathbf{e}_{\text{one}})^{-1} \right\}.$$

However,

$$\mathbf{e}_{\text{one}}^T \Xi^{-1} \mathbf{e}_{\text{one}} = \gamma^T \left\{ (\mathbf{C} - \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T)^{-1} + (\beta^2 \sigma_u^2 / c_F) \Sigma_v^{-1} \right\}^{-1} \gamma,$$

and hence

$$n_p^{1/2} (\hat{\beta}_{\text{ts}} - \beta) \rightarrow \text{Normal} \left\{ 0, \sigma_{\text{ts}}^2 = (\gamma^T \Omega \gamma)^{-1} \right\};$$

$$\Omega = \left\{ (\mathbf{C} - \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T)^{-1} + (\beta^2 \sigma_\varepsilon^2 / c_F) \Sigma_v^{-1} \right\}^{-1}. \quad (\text{A.5})$$

A.6. Asymptotic Equivalence of MLE and Two-Stage Methods

Here we show that the maximum likelihood and two-stage estimators of β are asymptotically equivalent. Write $\theta = c_F / \sigma_u^2$ and $\mathbf{N} = \mathbf{C} - \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T$. Studying (A.3) and (A.4), we have to show that $\Omega = \Lambda$ where for completeness

$$\Omega = \left\{ (\mathbf{C} - \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T)^{-1} + (\beta^2 \sigma_\varepsilon^2 / c_F) \Sigma_v^{-1} \right\}^{-1} = \left\{ (\mathbf{C} - \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T)^{-1} + (\beta^2 / \theta) \Sigma_v^{-1} \right\}^{-1};$$

$$\Lambda = \mathbf{C} - \begin{bmatrix} \mathbf{B} & \beta \mathbf{C} \end{bmatrix} \begin{bmatrix} \mathbf{A} & \beta \mathbf{B}^T \\ \beta \mathbf{B} & \theta \Sigma_v + \beta^2 \mathbf{C} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{B}^T \\ \beta \mathbf{C} \end{bmatrix}.$$

$$\mathbf{V} = \begin{bmatrix} A & \beta \mathbf{B}^T \\ \beta \mathbf{B} & \theta \Sigma_v + \beta^2 \mathbf{C} \end{bmatrix}.$$

Then

$$\mathbf{V}^{-1} = \begin{bmatrix} \mathbf{a}_{11} & \mathbf{a}_{12} \\ \mathbf{a}_{21} & \mathbf{a}_{22} \end{bmatrix},$$

where

$$\begin{aligned} \mathbf{a}_{22} &= (\theta \Sigma_v + \beta^2 \mathbf{N})^{-1} = \theta^{-1} \{ \Sigma_v + (\beta^2 / \theta) \mathbf{N} \}^{-1} = \theta^{-1} \left[\Sigma_v \left\{ \mathbf{N}^{-1} + (\beta^2 / \theta) \Sigma_v^{-1} \right\} \mathbf{N} \right]^{-1} = \theta^{-1} \mathbf{N}^{-1} \boldsymbol{\Omega} \Sigma_v^{-1}, \\ a_{11} &= A^{-1} \{ 1 + \beta^2 \mathbf{B}^T \mathbf{a}_{22} \mathbf{B} A^{-1} \}, \quad a_{12} = -\beta A^{-1} \mathbf{B}^T \mathbf{a}_{22} \text{ and } a_{21} = -\beta \mathbf{a}_{22} \mathbf{B} A^{-1}. \end{aligned}$$

$$\begin{aligned} \boldsymbol{\Lambda} &= \mathbf{C} - \mathbf{B} \mathbf{a}_{11} \mathbf{B}^T - \beta \mathbf{C} \mathbf{a}_{21} \mathbf{B}^T - \beta \mathbf{B} \mathbf{a}_{12} \mathbf{C} - \beta^2 \mathbf{C} \mathbf{a}_{22} \mathbf{C} \\ &= \mathbf{C} - \mathbf{B} A^{-1} \mathbf{B}^T - \beta^2 \mathbf{B} A^{-1} \mathbf{B}^T \mathbf{a}_{22} \mathbf{B} A^{-1} \mathbf{B}^T + \beta^2 \mathbf{C} \mathbf{a}_{22} \mathbf{B} A^{-1} \mathbf{B}^T + \beta^2 \mathbf{B} A^{-1} \mathbf{B}^T \mathbf{a}_{22} \mathbf{C} - \beta^2 \mathbf{C} \mathbf{a}_{22} \mathbf{C} \\ &= \mathbf{N} + \beta^2 \mathbf{N} \mathbf{a}_{22} \mathbf{B} A^{-1} \mathbf{B}^T = \beta^2 \mathbf{N} \mathbf{a}_{22} \mathbf{C} \\ &= \mathbf{N} - \beta^2 \mathbf{N} \mathbf{a}_{22} \mathbf{N} \\ &= \mathbf{N} - (\beta^2 / \theta) \boldsymbol{\Omega} \Sigma_v^{-1} \mathbf{N} \\ &= (\beta^2 / \theta) \left\{ (\theta / \beta^2) \mathbf{I} - \boldsymbol{\Omega} \Sigma_v^{-1} \right\} \mathbf{N} \\ &= (\beta^2 / \theta) \boldsymbol{\Omega} \left\{ (\theta / \beta^2) \boldsymbol{\Omega}^{-1} - \Sigma_v^{-1} \right\} \mathbf{N} \\ &= \boldsymbol{\Omega} \mathbf{N}^{-1} \mathbf{N} \\ &= \boldsymbol{\Omega} \end{aligned}$$

so that the MLE and the two-stage estimator are asymptotically equivalent. This completes the argument.

A.7. Asymptotic Non-Efficiency of Standard Regression Calibration

Here we show that the regression calibration estimator is not generally asymptotically equivalent to the other estimators. To do this, we simply need to exhibit a special case, and the simplest such case arises when $\Sigma_v = \mathbf{I}$ and $\mathbf{B} = 0$. When $\mathbf{B} = 0$, for the MLE,

$$\mathfrak{J} = \begin{bmatrix} A & 0 & 0 \\ 0 & (c_F / \sigma_\varepsilon^2) \mathbf{I} + \beta^2 \mathbf{C} & \beta \mathbf{C} \boldsymbol{\gamma} \\ 0 & \beta \boldsymbol{\gamma}^T \mathbf{C} & \boldsymbol{\gamma}^T \mathbf{C} \boldsymbol{\gamma} \end{bmatrix},$$

and hence

$$\sigma_{\text{MLE}}^2 = \mathbf{e}_1^T \mathfrak{J}^{-1} \mathbf{e}_1 = \left(\boldsymbol{\gamma}^T \left[\mathbf{C} - \beta^2 \mathbf{C} \{ \beta^2 \mathbf{C} + (c_F / \sigma_u^2) \mathbf{I} \}^{-1} \mathbf{C} \right] \boldsymbol{\gamma} \right)^{-1} = \sigma_{\text{TS}}^2,$$

the last following from our more general result.

For standard regression calibration, we get a different answer. Let $\mathbf{D}_*$ and $\mathbf{M}_*$ be the lower non-diagonal versions of $\mathbf{D}$ and $\mathbf{M}$, so that

$$\mathbf{D}_* = \begin{bmatrix} c_F \mathbf{I} & 0 \\ \beta \boldsymbol{\gamma}^T \mathbf{C} & \boldsymbol{\gamma}^T \mathbf{C} \boldsymbol{\gamma} \end{bmatrix}; \mathbf{M}_* = \begin{bmatrix} \sigma_\varepsilon^2 c_F \mathbf{I} & 0 \\ 0 & \boldsymbol{\gamma}^T \mathbf{C} \boldsymbol{\gamma} \end{bmatrix}.$$

Then

$$\mathbf{D}_*^{-1} = \begin{bmatrix} c_F^{-1} \mathbf{I} & 0 \\ -\{\beta / (c_F \boldsymbol{\gamma}^T \mathbf{C} \boldsymbol{\gamma})\} \boldsymbol{\gamma}^T \mathbf{C} & (\boldsymbol{\gamma}^T \mathbf{C} \boldsymbol{\gamma})^{-1} \end{bmatrix}.$$

Direct calculations show that

$$\sigma_{RC}^2 = (\boldsymbol{\gamma}^T \mathbf{C} \boldsymbol{\gamma})^{-1} \left[1 + \{\beta^2 \sigma_u^2 / (c_F \boldsymbol{\gamma}^T \mathbf{C} \boldsymbol{\gamma})\} \boldsymbol{\gamma}^T \mathbf{C} \mathbf{C} \boldsymbol{\gamma} \right].$$

This is in general different from the maximum likelihood and two-stage variance. It does coincide with them in the case of linear regression, where $\mathbf{C} = \mathbf{I} / \sigma_\varepsilon^2$, and in the case when $\beta = 0$.

Table A1

Parameters used in simulation study with lognormal $\mathbf{W}$.

Case	γ_1	γ_2	β_0	β_1	σ_ε^2	R^2 *	$\beta_1^2 \sigma_\varepsilon^2$	Odds Ratio $\exp(2\beta_1)$	Probability of Disease in Population
1	0.5	0.3	-3	0.50	0.51	0.49	0.13	2.7	0.05
2	0.4	0.2	-2	0.75	0.72	0.28	0.41	4.5	0.14
3	0.3	0.1	-2	1.00	0.87	0.13	0.87	7.4	0.16
4	0.2	0.1	-1	1.25	0.93	0.07	1.45	12.2	0.32
5	0.1	0.05	0	1.50	0.98	0.02	2.21	20.1	0.50

* $R^2 = 1 - \sigma_\varepsilon^2$ is the R-squared of the regression of X on (W_1, W_2) .

Table A2

Results of simulation study with lognormal $\mathbf{W}$ and *normal error* (ε). Means, standard deviations (SD) and root mean squared errors (RMSE) of the estimators of β_1 .

Case	β_1	$\beta_1^2 \sigma_\varepsilon^2$	Original Two-Stage Method			Modified Two-Stage Method		Regression Calibration Method	
1	0.50	0.13	Mean		0.490		0.491		0.493
			SD		0.022		0.022		0.022
			RMSE		0.024		0.024		0.023
2	0.75	0.41	Mean		0.692		0.697		0.701
			SD		0.043		0.042		0.043
			RMSE		0.072		0.068		0.065

Case	β_1	$\beta_1^2 \sigma_e^2$		Original Two-Stage Method	Modified Two-Stage Method	Regression Calibration Method
3	1.00	0.87	Mean	0.842	0.860	0.871
			SD	0.090	0.083	0.083
			RMSE	0.182	0.163	0.153
4	1.25	1.45	Mean	0.933	0.960	0.984
			SD	0.130	0.123	0.127
			RMSE	0.343	0.315	0.295
5	1.50	2.21	Mean	0.913	1.010	1.070
			SD	0.341	0.323	0.337
			RMSE	0.679	0.587	0.546

Table A3

Results of simulation study with lognormal $\mathbf{W}$ and *homoscedastic lognormal error* (ϵ). Means, standard deviations (SD) and root mean squared errors (RMSE) of the estimators of β_1 .

Case	β_1	$\beta_1^2 \sigma_e^2$		Original Two-Stage Method	Modified Two-Stage Method	Regression Calibration Method
1	0.50	0.13	Mean	0.485	0.486	0.487
			SD	0.022	0.022	0.022
			RMSE	0.027	0.026	0.026
2	0.75	0.41	Mean	0.684	0.689	0.693
			SD	0.046	0.043	0.043
			RMSE	0.080	0.075	0.071
3	1.00	0.87	Mean	0.838	0.854	0.864
			SD	0.093	0.083	0.081
			RMSE	0.187	0.168	0.158
4	1.25	1.45	Mean	1.004	1.033	1.056
			SD	0.153	0.140	0.139
			RMSE	0.290	0.258	0.239
5	1.50	2.21	Mean	1.086	1.184	1.260
			SD	0.377	0.359	0.380
			RMSE	0.560	0.478	0.449

Table A4

Results of simulation study with lognormal $\mathbf{W}$ and *heteroscedastic lognormal error* (ϵ). Means, standard deviations (SD) and root mean squared errors (RMSE) of the estimators of β_1 .

Case	β_1	$\beta_1^2 \sigma_e^2$		Original Two-Stage Method	Modified Two-Stage Method	Regression Calibration Method
1	0.50	0.13	Mean	0.458	0.473	0.479
			SD	0.116	0.115	0.109

Case	β_1	$\beta_1^2 \sigma_\varepsilon^2$		Original Two-Stage Method	Modified Two-Stage Method	Regression Calibration Method
2	0.75	0.41	RMSE	0.123	0.118	0.111
			Mean	0.566	0.602	0.616
			SD	0.187	0.196	0.180
3	1.00	0.87	RMSE	0.262	0.246	0.224
			Mean	0.700	0.757	0.783
			SD	0.293	0.298	0.292
4	1.25	1.45	RMSE	0.419	0.385	0.364
			Mean	0.733	0.793	0.825
			SD	0.356	0.358	0.355
5	1.50	2.21	RMSE	0.628	0.581	0.554
			Mean	0.780	0.872	0.911
			SD	0.450	0.462	0.466
			RMSE	0.849	0.780	0.751

Appendix B: Asymptotic Theory for Logistic Regression

The purpose of this appendix is to provide a precise asymptotic analysis in the case of logistic regression, clearly stating where the regression calibration is used. In cases that the regression calibration is used, we will replace $=$ by $\doteq$.

Let $H(x) = \exp(x) / \{1 + \exp(x)\}$ be the logistic distribution function. Our model is that in the validation and primary studies, which are of size n_p and n_v , respectively,

$$\begin{aligned}
 X_{iv} &= \mathbf{W}_{iv}^T \boldsymbol{\gamma} + \varepsilon_{iv}; \\
 X_{ip} &= \mathbf{W}_{ip}^T \boldsymbol{\gamma} + \varepsilon_{ip}; \\
 0 &= E(\varepsilon_{iv}) = E(\varepsilon_{ip}); \\
 \sigma_u^2 &= \text{var}(\varepsilon_{iv}) = \text{var}(\varepsilon_{ip}); \\
 0 &= E(\mathbf{W}_{iv}); \\
 \Sigma_v &= \text{cov}(\mathbf{W}_{iv}); \\
 \text{pr}(Y_{ip} = 1 | X_{ip}) &= H(\alpha + X_{ip}\beta).
 \end{aligned}$$

We define $c_F = n_v/n_p$ and use the fact that $n_p^{1/2}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \rightarrow \text{Normal}\{\mathbf{0}, (\sigma_\varepsilon^2/c_F) \sum_v^{-1}\}$.

B.1. Two-stage method

B.1.1. Basic Theory

The two-stage method makes the assumption that there is a unique $(\alpha_0, \boldsymbol{\alpha}_1)$ such that

$$0 = E \left[(1, \mathbf{W}_{ip}^T)^T \{Y - H(\alpha_0 + \mathbf{W}_{ip}^T \boldsymbol{\alpha}_1)\} \right]. \quad (\text{B.1})$$

In this case, $(\alpha_0, \boldsymbol{\alpha}_1)$ are the parameters that logistic regression of Y on $\mathbf{W}$ in the primary study actually converges to in probability. It should be noted that, precisely,

$$\text{pr}(Y_{ip}=1|\mathbf{W}_{ip}) \neq H(\alpha_0 + \mathbf{W}_{ip}^T \boldsymbol{\alpha}_1). \quad (\text{B.2})$$

In other word, (B.2) says that the observed data do not follow a logistic regression exactly. Logistic regression of Y on $\mathbf{W}$ in the primary study means solving the estimating equation

$$0 = \sum_{i=1}^{n_p} (1, \mathbf{W}_{ip}^T)^T \{Y_{ip} - H(\widehat{\alpha}_0 + \mathbf{W}_{ip}^T \widehat{\boldsymbol{\alpha}}_1)\}.$$

Define $v_{ip} = Y_{ip} - H(\alpha_0 + \mathbf{W}_{ip}^T \boldsymbol{\alpha}_1)$, and make the further definitions

$$\begin{aligned} \mathbf{Q}_1 &= E \left\{ (1, \mathbf{W}_{ip}^T)^T (1, \mathbf{W}_{ip}^T) v_{ip}^2 \right\}; \\ \mathbf{Q}_2 &= E \left\{ (1, \mathbf{W}_{ip}^T)^T (1, \mathbf{W}_{ip}^T) H^{(1)}(\alpha_0 + \mathbf{W}_{ip}^T \boldsymbol{\alpha}_1) \right\}, \end{aligned}$$

where $H^{(1)}(x) = H(x)\{1 - H(x)\}$. It is a consequence of standard estimating equation theory that

$$n_p^{1/2}(\widehat{\alpha}_0 - \alpha_0, \widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1)^T \rightarrow \text{Normal}(\mathbf{0}, \mathbf{Q}_2^{-1} \mathbf{Q}_1 \mathbf{Q}_2^{-1}).$$

This means that

$$n_p^{1/2}(\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1)^T \rightarrow \text{Normal}(\mathbf{0}, \mathbf{Q}_3 = [\mathbf{0}, \mathbf{I}] \mathbf{Q}_2^{-1} \mathbf{Q}_1 \mathbf{Q}_2^{-1} [\mathbf{0}, \mathbf{I}]^T).$$

The two-stage method considers a weighted linear combination of $\{\text{diag}(\boldsymbol{\gamma})\}^{-1} \widehat{\boldsymbol{\alpha}}_1$. We have that

$$\begin{aligned} n_p^{1/2} \left[\{\text{diag}(\widehat{\boldsymbol{\gamma}})\}^{-1} \widehat{\boldsymbol{\alpha}}_1 - \{\text{diag}(\boldsymbol{\gamma})\}^{-1} \boldsymbol{\alpha}_1 \right] &= \{\text{diag}(\boldsymbol{\gamma})\}^{-1} n_p^{1/2}(\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1) + n_p^{1/2} \left[\{\text{diag}(\widehat{\boldsymbol{\gamma}})\}^{-1} - \{\text{diag}(\boldsymbol{\gamma})\}^{-1} \right] \boldsymbol{\alpha}_1 + o_p(1) \\ &= \{\text{diag}(\boldsymbol{\gamma})\}^{-1} \left[n_p^{1/2}(\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1) + \text{diag}(\boldsymbol{\alpha}_1) \{\text{diag}(\boldsymbol{\gamma})\}^{-1} n_p^{1/2}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \right] + o_p(1) \\ &\rightarrow \text{Normal}(\mathbf{0}, \mathbf{Q}_4), \end{aligned}$$

where

$$\mathbf{Q}_4 = \{\text{diag}(\boldsymbol{\gamma})\}^{-1} \left[\mathbf{Q}_3 + \text{diag}(\boldsymbol{\alpha}_1) \{\text{diag}(\boldsymbol{\gamma})\}^{-1} (\sigma_u^2 / c_f) \sum_v^{-1} \text{diag}(\boldsymbol{\alpha}_1) \{\text{diag}(\boldsymbol{\gamma})\}^{-1} \right] \{\text{diag}(\boldsymbol{\gamma})\}^{-1}.$$

If $\mathbf{e}_{\text{one}} = (1, 1, \dots, 1)^T$, then the optimal weighted combination in terms of variance of $\{\text{diag}(\widehat{\boldsymbol{\gamma}})\}^{-1} \widehat{\boldsymbol{\alpha}}_1$ has weights $\boldsymbol{\omega} = \mathbf{Q}_4^{-1} \mathbf{e}_{\text{one}} (\mathbf{e}_{\text{one}}^T \mathbf{Q}_4^{-1} \mathbf{e}_{\text{one}})^{-1}$. The two stage estimator of $\boldsymbol{\beta}$ then is $\widehat{\boldsymbol{\beta}}_{\text{TS}} = \boldsymbol{\omega}^T \{\text{diag}(\boldsymbol{\gamma})\}^{-1} \widehat{\boldsymbol{\alpha}}_1$, and we have that

$$n_p^{1/2}(\widehat{\beta}_{TS} - \omega^T \{\text{diag}(\gamma)\}^{-1} \alpha_1) \rightarrow \text{Normal}\{0, (\mathbf{e}_{\text{one}}^T \mathbf{Q}_4^{-1} \mathbf{e}_{\text{one}})^{-1}\}. \quad (\text{B.3})$$

Of course, an implication of (B.3) is that the two-stage estimator converges to $\omega^T \{\text{diag}(\gamma)\}^{-1} \alpha_1$.

B.1.2. Implications of the regression calibration approximation

Equation (B.3) makes it clear that the two-stage method actually converges to $\omega^T \{\text{diag}(\gamma)\}^{-1} \alpha_1$. However, $\omega = \omega(\mathbf{Q}_1, \mathbf{Q}_2, \gamma, \alpha_1)$ depends on $(\mathbf{Q}_1, \mathbf{Q}_2, \gamma, \alpha_1)$. Since (α_0, α_1) are the solutions to the integral equations (B.1), which in turn depend on the distributions of $(X, \mathbf{W})$ in the primary study, the asymptotic variance (B.4) is not particularly useful for comparison with regression calibration.

However, there are important simplifications if one makes the same regression calibration approximations that are made by Weller, et al.[1], namely that for some β_{TS} ,

$$\text{pr}(Y_{ip}=1|\mathbf{W}_{ip}) \doteq H(\alpha_0 + \mathbf{W}_{ip}^T \alpha_1); \quad (\text{B.4})$$

$$\beta_{TS} \mathbf{e} \doteq \{\text{diag}(\gamma)\}^{-1} \alpha_1, \quad (\text{B.5})$$

where in both cases we have invoked the term $\doteq$ to reflect the regression calibration approximation. In this case, it is readily seen from (B.4) that $\mathbf{Q}_1 \doteq \mathbf{Q}_2$, $\omega^T \{\text{diag}(\gamma)\}^{-1} \alpha_1 \doteq \beta_{TS}$, and then the regression calibration approximation to the asymptotic variance of the two-stage method reduces to the result in the Appendix A. In particular, define

$V_{ip} = H^{(1)}(\alpha_0 + \mathbf{W}_{ip}^T \alpha_1)$. Then we have that $\mathbf{Q}_3 \doteq (\mathbf{C} - \mathbf{B} \mathbf{A}^{-1} \mathbf{B}^T)^{-1}$, which agrees with the result in Appendix A. Using (B.5), we also have

$$\mathbf{Q}_4 = \{\text{diag}(\gamma)\}^{-1} \left\{ (\mathbf{C} - \mathbf{B} \mathbf{A}^{-1} \mathbf{B}^T)^{-1} + (\beta_{TS}^2 \sigma_u^2 / c_F) \sum_v^{-1} \right\} \{\text{diag}(\gamma)\}^{-1},$$

which also agrees with the results in Appendix A.

B.2. Regression calibration method

We will distinguish here between the regression calibration method and the regression calibration approximation.

B.2.1. Basic theory

The regression calibration method makes the assumption that there is a unique (β_0, β_{RC}) such that

$$0 = E \left[(1, \mathbf{W}_{ip}^T \gamma)^T \{Y_{ip} - H(\beta_0 + \mathbf{W}_{ip}^T \gamma \beta_{RC})\} \right]. \quad (\text{B.6})$$

Then $(\hat{\beta}_0, \hat{\beta}_1)$ solve

$$0 = \sum_{i=1}^{n_p} (1, \mathbf{W}_{ip}^T \hat{\boldsymbol{\gamma}})^T \{Y_{ip} - H(\hat{\beta}_0 + \mathbf{W}_{ip}^T \hat{\boldsymbol{\gamma}} \hat{\beta}_{RC})\}. \quad (\text{B.7})$$

This means that the estimating equations are

$$\begin{aligned} 0 = & n_p^{-1/2} \sum_{i=1}^{n_v} \{0, \mathbf{W}_{iv}^T (X_{iv} - \mathbf{W}_{iv}^T \hat{\boldsymbol{\gamma}}), 0\}^T + n_p^{-1/2} \sum_{i=1}^{n_p} \begin{bmatrix} = Y_{ip} - H(\hat{\beta}_0 + \mathbf{W}_{ip}^T \hat{\boldsymbol{\gamma}} \hat{\beta}_{RC}) \\ 0 \\ \mathbf{W}_{ip}^T \hat{\boldsymbol{\gamma}} \{Y_{ip} - H(\hat{\beta}_0 + \mathbf{W}_{ip}^T \hat{\boldsymbol{\gamma}} \hat{\beta}_{RC})\} \end{bmatrix} \\ & - n_p^{-1/2} \begin{bmatrix} \sum_{i=1}^{n_p} \{Y_{ip} - H(\beta_0 + \mathbf{W}_{ip}^T \boldsymbol{\gamma} \beta_{RC})\} \\ \sum_{i=1}^{n_v} \mathbf{W}_{iv}^T (X_{iv} - \mathbf{W}_{iv}^T \boldsymbol{\gamma}) \\ \sum_{i=1}^{n_p} \mathbf{W}_{ip}^T \boldsymbol{\gamma} \{Y_{ip} - H(\beta_0 + \mathbf{W}_{ip}^T \boldsymbol{\gamma} \beta_{RC})\} \end{bmatrix} - \mathbf{D}_1 \begin{bmatrix} n_p^{1/2} (\hat{\beta}_0 - \beta_0) \\ n_p^{1/2} (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \\ n_p^{1/2} (\hat{\beta}_{RC} - \beta_{RC}) \end{bmatrix} + o_p(1), \end{aligned}$$

where if $V_{ip} = H^{(1)}(\beta_0 + \mathbf{W}_{ip}^T \boldsymbol{\gamma} \beta_{RC})$,

$$\mathbf{D}_1 = E \begin{bmatrix} V_{ip} & \mathbf{W}_{ip}^T V_{ip} \beta_{RC} & \mathbf{W}_{ip}^T \boldsymbol{\gamma} V_{ip} \\ 0 & c_F \sum_v & 0 \\ \mathbf{W}_{ip}^T \boldsymbol{\gamma} V_{ip} & \boldsymbol{\gamma}^T \mathbf{W}_{ip} \mathbf{W}_{ip}^T V_{ip} \beta_{RC} - \mathbf{W}_{ip}^T \{Y_{ip} - H(\beta_0 + \mathbf{W}_{ip}^T \boldsymbol{\gamma} \beta_{RC})\} & \boldsymbol{\gamma}^T \mathbf{W}_{ip} \mathbf{W}_{ip}^T \boldsymbol{\gamma} V_{ip} \end{bmatrix}.$$

Letting $v_{ip} = Y_{ip} - H(\beta_0 + \mathbf{W}_{ip}^T \boldsymbol{\gamma} \beta_{RC})$, and defining

$$\mathbf{D}_2 = E \begin{bmatrix} v_{ip}^2 & 0 & \mathbf{W}_{ip}^T \boldsymbol{\gamma} v_{ip}^2 \\ 0 & c_F \sigma_\varepsilon^2 \sum_v & 0 \\ \mathbf{W}_{ip}^T \boldsymbol{\gamma} v_{ip}^2 & 0 & (\mathbf{W}_{ip}^T \boldsymbol{\gamma})^2 v_{ip}^2 \end{bmatrix},$$

it follows that if $\mathbf{e}_1 = (0, 0, \dots, 0, 1)^T$,

$$n_p^{1/2} (\hat{\beta}_{RC} - \beta_{RC}) \rightarrow \text{Normal}\{0, \mathbf{D}_3 = \mathbf{e}_1^T \mathbf{D}_1^{-1} \mathbf{D}_2 (\mathbf{D}_1^{-1})^T \mathbf{e}_1\}. \quad (\text{B.8})$$

B.2.2. Implications of the regression calibration approximation

Like (B.3), equation (B.8) is not particularly useful for making comparisons among various methods because of the need to solve the integral equation (B.6) that in turn depends on the distribution of $\mathbf{W}_{ip}$ etc. However, the equivalent regression calibration approximation to (B.4)–(B.5) is that

$$\text{pr}(Y_{ip}=1|\mathbf{W}_{ip}) \doteq H(\beta_0 + \mathbf{W}_{ip}^T \boldsymbol{\gamma} \beta_{RC}). \quad (\text{B.9})$$

Under (B.4)–(B.5) and (B.9), $\beta_{TS} \doteq \beta_{RC}$. Equation (B.9) implies that

$E(v_{ip}^2|\mathbf{W}_{ip}) \doteq E(V_{ip}|\mathbf{W}_{ip})$. In this case, $\mathbf{D}_1$ and $\mathbf{D}_2$ simplify to

$$\mathbf{D}_1 \doteq E \begin{bmatrix} A & \mathbf{B}^T \beta_{RC} & \mathbf{B}^T \gamma \\ 0 & c_F \sum_v & 0 \\ \mathbf{B}^T \gamma & \gamma^T \mathbf{C} \beta_{RC} & \gamma^T \mathbf{C} \gamma \end{bmatrix};$$

$$\mathbf{D}_2 \doteq E \begin{bmatrix} A & 0 & \mathbf{B}^T \gamma \\ 0 & c_F \sigma_\varepsilon^2 \sum_v & \mathbf{B}^T \gamma \\ \mathbf{B}^T \gamma & 0 & \gamma^T \mathbf{C} \gamma \end{bmatrix},$$

exactly as described in Appendix A.

Appendix C: Simulations with Lognormal W

In this appendix, we present results of three sets of simulations where surrogates W_1 and W_2 had bivariate lognormal distributions. We first generated (Z_1, Z_2) as bivariate standard normal with correlation ρ_* . Let $\sigma = \sqrt{\log(5)}$, $\mu_* = \exp(\sigma^2/2)$ and $\sigma_*^2 = \exp(2\sigma^2) - \exp(\sigma^2)$, so that the coefficient of variation is $\sigma_*/\mu_* = 2$. By simulation, we found ρ_* such that $\text{corr}\{\exp(\sigma Z_1), \exp(\sigma Z_2)\} = \rho$. Then (W_1, W_2) defined by $W_j = \{\exp(\sigma Z_j) - \mu_*\}/\sigma_*$, $j=1,2$ has the desired bivariate lognormal distribution with mean zero, variance one, and correlation ρ .

As with other simulations described in the main text, we generated X as $X = \gamma_0 + \gamma_1 W_1 + \gamma_2 W_2 + \varepsilon$, where ε was uncorrelated with (W_1, W_2) and $E(\varepsilon) = 0$, $\sigma_\varepsilon^2 = 1 - (\gamma_1^2 + \gamma_2^2 + 2\rho\gamma_1\gamma_2)$, thus ensuring $E(X) = 0$, $\text{var}(X) = 1$. As before, we generated Y as a binary variable with $\text{pr}(Y=1/X) = H(\beta_0 + \beta_1 X)$, where $H(\cdot)$ is the logistic distribution function. Then we set the true logistic intercept β_0 and the true logistic slope β_1 so that the odds ratio for increasing X by two standard deviations of its distribution was equal to a predefined value OR, and the probability of disease in the source population was equal to a predefined value π .

In the first set of simulations, we generated ε to be normal and independent of (W_1, W_2) . In the second set, we set $\varepsilon = \delta - E(\delta)$, where δ was independent of (W_1, W_2) and had a

lognormal distribution with $E(\delta) = \sqrt{1 - (\gamma_1^2 + \gamma_2^2 + 2\rho\gamma_1\gamma_2)}$ and $CV = 1$. In the third set of simulations, we considered a case of heteroscedastic error $\varepsilon = (\gamma_1 W_1 + \gamma_2 W_2) \times (\delta - 1)$, δ was independent of (W_1, W_2) , lognormal with mean $E(\delta) = 1$, variance $\text{var}(\delta) = \{1 - (\gamma_1^2 + \gamma_2^2 + 2\rho\gamma_1\gamma_2)\} / (\gamma_1^2 + \gamma_2^2 + 2\rho\gamma_1\gamma_2)$, and $CV(\delta) = \text{var}(\delta)$.

In all simulations, we set $\rho = 0.5$, $\gamma_0 = 0$, $n_v = 1,000$ and $n_p = 100,000$. Other parameters for each simulation are given in Table A1. Tables A2 – A4 present the results for simulation sets 1 to 3, respectively.

Figure 1. Box plots of five randomly generated lognormal data sets of size 10,000 with coefficient of variation = 1.0.

Figure 2.
Box plots of five randomly generated gamma data sets of size 10,000 with coefficient of variation = 1.0.

Table 1Parameters used in simulation study with normal **W**.

Case	γ_1	γ_2	β_0	β_1	σ_ε^2	R^2 *	$\beta_1^2 \sigma_\varepsilon^2$	Odds Ratio $\exp(2\beta_1)$	Probability of Disease in Population
1	0.5	0.3	-3	0.50	0.51	0.49	0.13	2.7	0.05
2	0.4	0.2	-2	0.75	0.72	0.28	0.41	4.5	0.14
3	0.3	0.1	-2	1.00	0.87	0.13	0.87	7.4	0.16
4	0.2	0.1	-1	1.25	0.93	0.07	1.45	12.2	0.32
5	0.1	0.05	0	1.50	0.98	0.02	2.21	20.1	0.50

* $R^2 = 1 - \sigma_\varepsilon^2$ is the R-squared of the regression of X on (W_1, W_2).

Table 2

Results of simulation study with normal W. Means, standard deviations (SD) and root mean squared errors (RMSE) of the estimators of β_1 .

Case	β_1	$\beta_1^2 \sigma_\varepsilon^2$	Original Two-Stage Method			Modified Two-Stage Method			Regression Calibration Method		
1	0.50	0.13	Mean	0.494		0.495			0.497		
			SD	0.067		0.067			0.067		
			RMSE	0.067		0.067			0.067		
2	0.75	0.41	Mean	0.706		0.710			0.714		
			SD	0.067		0.066			0.067		
			RMSE	0.080		0.077			0.076		
3	1.00	0.87	Mean	0.867		0.882			0.894		
			SD	0.111		0.108			0.108		
			RMSE	0.173		0.160			0.151		
4	1.25	1.45	Mean	0.937		0.964			0.987		
			SD	0.150		0.143			0.143		
			RMSE	0.347		0.320			0.299		
5	1.50	2.21	Mean	0.914		1.002			1.072		
			SD	0.336		0.326			0.341		
			RMSE	0.675		0.595			0.547		

Table 3

Parameters used in simulation study with categorical W.

Case	γ_1	γ_2	β_0	β_1	σ_ε^2	R^2	$\beta_1^2 \sigma_\varepsilon^2$	Odds Ratio $\exp(2\beta)_1$	Probability of Disease in Population
1	1.5	1.0	-3	0.50	0.59	0.41	0.15	2.7	0.05
2	1.5	0.5	-2	0.75	0.61	0.39	0.22	4.5	0.14
3	1.0	0.5	-2	1.00	0.83	0.17	0.83	7.4	0.16
4	1.0	0.25	-1	1.25	0.82	0.18	1.28	12.2	0.32
5	0.5	0.25	0	1.50	0.96	0.04	2.16	20.1	0.50

Table 4

Results of simulation study with categorical W. Means, standard deviations (SD) and root mean squared errors (RMSE) of the estimators of β_1 .

Case	β_1	$\beta_1^2 \sigma_\varepsilon^2$	Original Two-Stage Method			Modified Two-Stage Method			Regression Calibration Method		
1	0.50	0.15	Mean	0.497		0.495		0.497			0.497
			SD	0.077		0.077		0.077			0.077
			RMSE	0.077		0.077		0.077			0.077
2	0.75	0.22	Mean	0.718		0.715		0.718			0.718
			SD	0.054		0.053		0.054			0.054
			RMSE	0.063		0.064		0.063			0.063
3	1.00	0.83	Mean	0.898		0.901		0.901			0.901
			SD	0.093		0.093		0.094			0.094
			RMSE	0.138		0.136		0.137			0.137
4	1.25	1.28	Mean	1.008		0.997		1.007			1.007
			SD	0.087		0.086		0.086			0.086
			RMSE	0.257		0.267		0.258			0.258
5	1.50	2.16	Mean	1.053		1.031		1.074			1.074
			SD	0.197		0.199		0.195			0.195
			RMSE	0.488		0.509		0.469			0.469