Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Dane R. Van Domelen Date

Measurement error methods for unmeasured confounding and pooling

By

Dane R. Van Domelen
Doctor of Philosophy

Biostatistics

Robert H. Lyles
Advisor

Yijian Huang
Committee Member

Amita K. Manatunga
Committee Member

Emily M. Mitchell
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Measurement error methods for unmeasured confounding and pooling

By

Dane R. Van Domelen
M.Sc., Emory University, 2016

Advisor: Robert H. Lyles, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2018

Abstract

Measurement error methods for unmeasured confounding and pooling
By Dane R. Van Domelen

Epidemiologists increasingly utilize existing datasets to explore exposure-disease relationships. A common problem is that one or more covariates may not be available. In Chapter 1, we compare methods for handling unmeasured confounding when validation data can be obtained. We consider propensity score calibration as well as maximum likelihood and regression calibration from the measurement error literature, both of which require specifying a model for the unmeasured confounder given exposure, disease model covariates, and perhaps additional covariates. We apply the methods to assess whether low Vitamin D is associated with fecundity controlling for age, overweight status, and caloric intake, by combining a primary dataset missing caloric intake with a smaller validation dataset. We propose several modifications to propensity score calibration to relax a critical surrogacy assumption, leading to improved performance but nullifying an appealing identifiability property of the original method.

In the logistic regression setting, measuring biomarkers in combined samples ("pools") from multiple cases or controls can lead to large gains in statistical efficiency. Two types of error threaten validity: assay-related measurement error, and processing error caused by forming pools. In Chapter 2, we present a likelihood approach to correct for both errors. We assume the biomarker level given covariates is normally distributed, and measurement and processing errors are independent, normally distributed, and not dependent on pool size. Our approach accommodates replicate measurements, which are not required for identifiability but improve stability. We apply our methods to a reproductive health dataset with pools of size 1 and 2 and replicates and assess validity and efficiency via simulations.

In Chapter 3, we present a logistic regression approach and a discriminant function approach for estimating the covariate-adjusted odds ratio relating a binary outcome to a right-skewed biomarker measured in homogeneous pools. Both assume multiplicative lognormal (rather than additive normal) measurement and processing errors acting on the poolwise mean and utilize constant-scale Gamma models for the biomarker level. In the motivating example, AIC favors these models over their normal counterparts from Chapter 2, although substantive results are similar. Our methods are implemented in the R package **pooling**.

Measurement error methods for unmeasured confounding and pooling

By

Dane R. Van Domelen
M.Sc., Emory University, 2016

Advisor: Robert H. Lyles, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2018

Acknowledgments

This research was supported by the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-0940903. The views expressed in this dissertation are those of the authors, and no official endorsement by the Department of Health and Human Services, or the Agency for Healthcare Research and Quality, or the National Science Foundation, is intended or should be inferred.

Contents

List of Tables

List of Figures

Chapter 1: Measurement error methods in the unmeasured confounding setting

1.1 Introduction

We consider the situation in which an investigator wishes to fit a generalized linear model to estimate the association between an exposure X and an outcome Y adjusted for covariates $(\boldsymbol{Z}, \boldsymbol{C})$:

$$g[E(Y)] = \beta_0 + \beta_x X + \boldsymbol{\beta}_{\boldsymbol{z}}^T \boldsymbol{Z} + \boldsymbol{\beta}_{\boldsymbol{c}}^T \boldsymbol{C} \tag{1.1}$$

but the dataset of interest is missing $\boldsymbol{Z}$. This is a very common problem in observational research, as epidemiologists often utilize existing data from large population-based studies or disease registries to assess exposure-disease relationships (Smith *et al.*, 2011). Perhaps more often than not, the most relevant existing dataset to explore a particular research question does not include data on every single variable of interest.

In such a scenario, the simplest approach is to fit Eq. 1.1 without $\boldsymbol{Z}$ and obtain what might be termed "naive" estimates $\hat{\boldsymbol{\beta}}^* = (\hat{\beta}_x^*, \hat{\boldsymbol{\beta}}_{\boldsymbol{c}}^{*T})^T$. These estimates are generally biased for $\boldsymbol{\beta} = (\beta_x, \boldsymbol{\beta}_{\boldsymbol{c}}^T)^T$, but may be informative in cases where the direction of confounding due

to $\boldsymbol{Z}$ can be determined. For example, if $\hat{\beta}_x^*$ is statistically and clinically significant and is identified as a lower bound for β_x, then the effect of X on Y controlling for $\boldsymbol{C}$ and $\boldsymbol{Z}$ is also statistically and clinically significant (VanderWeele *et al.*, 2008).

In many cases, one might prefer an unbiased and consistent method for direct estimation of $\boldsymbol{\beta}$. When validation data on $(X, \boldsymbol{Z}, \boldsymbol{C})$ are available or feasible to obtain, corrective methods from the measurement error literature could be implemented.

One could employ variants of regression calibration (RC) used, for example, by Lyles and Kupper (1997); Weller *et al.* (2007); Kipnis *et al.* (2012); Lyles and Kupper (2013) for handling covariate measurement error. These methods require specifying a model for the expected value of $\boldsymbol{Z}$ given $(X, \boldsymbol{C})$ and perhaps additional covariates $\boldsymbol{D}$ available in both the main study and validation study. The $\boldsymbol{D}$ here should not be confused with $D =$ disease status often used in epidemiology; we follow the $(\boldsymbol{C}, \boldsymbol{D})$ notation of Lyles and Kupper (2013).

In the case where Z is scalar and continuous, validation data may support a linear regression $Z|(X, \boldsymbol{D}, \boldsymbol{C}) \sim (\alpha_0 + \alpha_x X + \boldsymbol{\alpha_d^T D} + \boldsymbol{\alpha_c^T C},\ \sigma_\delta^2)$. If Z is skewed, an alternative approach is to assume $Z|(X, \boldsymbol{D}, \boldsymbol{C}) \sim LN(\alpha_0 + \alpha_x X + \boldsymbol{\alpha_d^T D} + \boldsymbol{\alpha_c^T C},\ \sigma_\delta^2)$.

In the Lyles and Kupper (2013) application of RC, the linear regression is fit via ordinary least squares to obtain $\hat{\boldsymbol{\alpha}}$, and then Eq. 1.1 is fit with $E(Z|X, \boldsymbol{D}, \boldsymbol{C}; \hat{\boldsymbol{\alpha}})$ in place of the unobserved Z's. RC gives consistent estimates of $\boldsymbol{\beta}$ in linear regression (Carroll *et al.*, 2006), approximately consistent estimates in logistic regression under certain conditions (Rosner *et al.*, 1989; Kuha, 1994), and often performs well for other generalized linear models (Carroll *et al.*, 2006).

With validation data on hand, one might be comfortable fully specifying the distribution of $\boldsymbol{Z}|(X, \boldsymbol{D}, \boldsymbol{C})$, and performing a maximum likelihood (ML) analysis (Lyles and Kupper, 2013). For example, one could use a linear regression similar to RC, but with random normal errors. Compared to RC, a two-model ML approach has some advantages (efficiency, flexibility) and some disadvantages (less ease of implementation, potential numerical instability,

extra distributional assumption).

If primary interest is in β_x and X is a binary exposure, another option is propensity score calibration, a method developed by Stürmer *et al.* (2005) specifically to handle unmeasured confounding. Briefly, this would involve obtaining validation data and fitting a model for the probability of X given the main study covariates $\boldsymbol{C}$, a model for the probability of X given the full covariate vector $(\boldsymbol{Z}, \boldsymbol{C})$, and a linear model relating the two. In the main study, $\hat{P}(X|\boldsymbol{C})$ is calculated for each subject, mapped to $\hat{P}(X|\boldsymbol{Z}, \boldsymbol{C})$, and the model for Y is fit with X and $\hat{P}(X|\boldsymbol{Z}, \boldsymbol{C})$ as predictors. This method is computationally simple and generalizes nicely to the case of several unmeasured confounders, but its validity depends critically on a surrogacy assumption that may be difficult to assess (see Section 1.2.5, pg. 14).

Measurement error methods like regression calibration and maximum likelihood have not typically been applied to the unmeasured confounding setting (Streeter *et al.*, 2017; Zhang *et al.*, 2018). This application seems natural. However, there are some aspects of the unmeasured confounding scenario that distinguish it from covariate measurement error. For example, imprecise versions of true predictors are not observed; in effect, there is no "measurement error." As a result, validation data is useful, while replication data is not. And because there are no imprecise versions of $\boldsymbol{Z}$, valid estimation via ML or RC often requires identifying variables that inform $\boldsymbol{Z}$ but not Y given $(X, \boldsymbol{Z}, \boldsymbol{C})$. Such variables are akin to instrumental variables used for causal inference (Greenland, 2000), but their purpose here is to provide identifiability rather than establish causality. Finally, in evaluating validity and efficiency, the regression coefficient for the error-prone variable is of primary interest in the measurement error setting, while a regression coefficient for a perfectly measured variable (X) would be of primary interest in the unmeasured confounding setting.

In this chapter, we consider the use of maximum likelihood, regression calibration, and propensity score calibration to correct for unmeasured confounding. Our motivating exam-

ple is estimation of the covariate-adjusted log-odds ratio relating low Vitamin D to incident pregnancy, using data from a clinical trial designed for a different purpose. A potentially important covariate, caloric intake, was not measured for any of the 995 main study subjects, but it was measured along with Vitamin D and other covariates in a separate study with 89 subjects. We apply the corrective methods to this motivating example, perform simulations modeled after the data to assess validity and efficiency of the methods, and provide recommendations for epidemiological research.

1.1.1 Confounding

For the most part, our focus is not on assessing causality or choosing an appropriate set of control variables to obtain an unconfounded effect estimate. We assume that investigators have already specified a regression model of interest, and the regression parameters represent quantities of epidemiological interest. Still, some background information on directed acyclic graphs, causality, and confounding is warranted.

Definition and consequences

Greenland *et al.* (1999) define confounding as the scenario in which the probability distribution for an outcome variable differs across levels of an exposure for reasons other than the effects of the exposure. The variables responsible are termed confounders.

Failure to adequately control for confounding can result in various incorrect conclusions, such as concluding that X affects Y when it truly does not, that X does not affect Y when it truly does, finding a protective or harmful effect when the opposite is true, or over or underestimating the true causal effect.

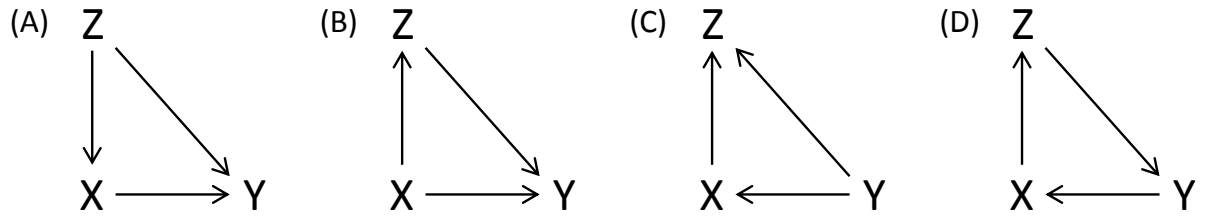

Figure 1.1: Four graphs showing assumed cause and effect relationships among variables.

Directed acyclic graphs

Directed acyclic graphs or DAGs are a useful graphical tool for conceptualizing epidemiological relationships and identifying confounders (Greenland *et al.*, 1999). Four simple graphs are shown in Figure 1.1 to illustrate important concepts.

The variables X, Z, and Y are termed nodes, and the lines connecting them are termed arcs or edges. Single-headed arrows reflect the direction of assumed cause and effect relationships. For example, Figure 1.1 (A) reflects the hypothesis that X has a causal effect on Y, and Z has a causal effect on both X and Y. Lack of a single-headed arrow from one variable to another implies the absence of a direct effect.

A path is any unbroken route that connects nodes. A directed or causal path is one in which each variable on the path causes the next. For example, in Figure 1.1 (A), the ZXY path is a causal path, while the XZY path is not. The latter path is termed a backdoor path form X to Y because it has an arrow pointing towards X.

The "directed" part of the term directed acyclic graph refers to the fact that all edges have a single or double headed arrow. The "acyclic" part refers to the fact that no directed path forms a closed loop. Figure 1.1 (D) is not a DAG because it has a cyclic path.

In Figure 1.1 (A), X has a direct effect on Y, or in other words X is a parent of Y and Y is a child of X. Z has both a direct effect on Y and an indirect effect through the ZXY path.

A variable is considered an ancestor of another if a directed path of arrows connects the

first to the second. For example, in Figure 1.1 (A), Z is an ancestor of both X and Y and X is an ancestor of Y. Similarly, X and Y are descendents of Z, and Y is also a descendent of X.

To illustrate colliding, in Figure 1.1 (C), the XZY path is said to be blocked by Z, or the path collides at Z, or Z is a collider on the path. Conversely, the XZY paths in Figure 1.1 (A) and (B) are unblocked because Z is not a collider in either case.

Two variables are marginally unassociated if there are no unblocked paths between the two variables. If the ZX edge in Figure 1.1 (A) was not there, then X and Z would be marginally unassociated, because the only path connecting them collides at Y.

An unblocked path must be either a directed path or a backdoor path through a shared ancestor. The former suggests a causal effect, while the latter suggests a confounded effect. A combination of the two can occur. For example, in Figure 1.1 (A), the unblocked XY path indicates a causal effect of X on Y, while the unblocked XZY path indicates a confounded association of X with Y. In other words, the association between X and Y is partially causal and partially confounded. A crude measure of association between X and Y represents the net result of both effects, which may be in the same or opposite directions, and may cancel each other entirely.

A causal effect of one variable on another is said to be mediated by a third variable if the causal path passes through that variable. For example, in Figure 1.1 (B), X has a direct effect on Y (the XY path), but also has an indirect effect on Y through the mediator Z (the XZY path).

DAGs: Assessing confounding

A common goal of epidemiological analysis is to estimate the total causal effect of an exposure on an outcome. The total causal effect may include direct and indirect effects, e.g. the XY direct effect and XZY indirect effect in Figure 1.1 (B). Once a DAG is constructed, one can

determine whether a crude estimate of the exposure effect is potentially confounded by one or several other variables.

After deleting all single-headed arrows coming from the exposure variable of interest, if there are no unblocked paths from the exposure to the outcome variable, then there is no potential confounding; otherwise, there is. This makes intuitive sense because an unblocked path from exposure to outcome after removing exposure effects means that the variables remain associated for reasons other than the exposure's effect.

In Figure 1.1 (A), the crude XY association is confounded, because when you remove the arrow emanating from X the unblocked backdoor path XZY still connects X to Y. In Figure 1.1 (B), the crude XY association is not confounded, because when you remove the arrows emanating from X there is no path that connects X to Y.

1.2 Methods

1.2.1 Notation

Using the nomenclature of Clayton *et al.* (1992), we refer to the regression model of interest as the true disease model (TDM), and the model for the unmeasured confounder as the measurement error model (MEM). The TDM specifies the relationship between the outcome Y and covariates $(X, Z, \boldsymbol{C}, \boldsymbol{B})$, where X is the exposure of interest, Z is a covariate missing in the primary dataset, $\boldsymbol{C}$ are covariates that are assumed to be related to Z, and $\boldsymbol{B}$ are covariates that are assumed to be unrelated to Z. The MEM specifies the relationship between Z and $(X, \boldsymbol{D}, \boldsymbol{C})$, which can be modeled using validation data. X could be absorbed into $\boldsymbol{C}$, but we leave it as separate to make it clear that the missing Z is a covariate rather than an exposure of primary interest in our scenario.

This setup differs from Weller *et al.* (2007) and Lyles and Kupper (2013) in the inclusion of covariates $\boldsymbol{B}$ which appear in the TDM but not in the MEM. Those references also focused on the measurement error problem, in which X is the dependent variable in the MEM. It is natural in that setting for all TDM covariates to also appear in the MEM, since covariates unrelated to X could simply be omitted from the TDM. In our scenario, the MEM outcome variable is Z; there could be covariates related to X and thus included in the TDM, but unrelated to Z. Leaving such variables out of the MEM might make it easier to find validation data, i.e. a dataset with $(X, Z, \boldsymbol{D}, \boldsymbol{C})$ rather than those variables plus $\boldsymbol{B}$.

1.2.2 Data types

We consider scenarios in which there is a main study and a validation study. Main study subjects are missing Z but have data on $(Y, X, \boldsymbol{D}, \boldsymbol{C}, \boldsymbol{B})$. In an internal validation study, Z is measured for a subset of main study subjects, such that all variables including the outcome are observed: $(Y, X, Z, \boldsymbol{D}, \boldsymbol{C}, \boldsymbol{B})$. In an external validation study, variables for fitting the MEM are observed in a different group of subjects: $(X, Z, \boldsymbol{D}, \boldsymbol{C})$ are available, but not Y or $\boldsymbol{B}$.

Using external validation data requires a transportability assumption (Carroll *et al.*, 2006), which means that the MEM applicable to variables in the validation study population is the same as that in the main study population. This assumption is typically unverifiable and may be suspect when demographics are quite different in the two studies.

1.2.3 Maximum likelihood

2-model ML

For main study subjects, $(Y, X, \boldsymbol{D}, \boldsymbol{C}, \boldsymbol{B})$ are observed, while Z is unobserved and has to be integrated out of the likelihood function. In the interest of specifying as few models as possible, a two-density factorization is:

$$\begin{aligned} L_i(\boldsymbol{\theta}) &= f(Y_i|X_i, \boldsymbol{D}_i, \boldsymbol{C}_i, \boldsymbol{B}_i) \\ &= \int_{Z_i} f(Y_i, Z_i|X_i, \boldsymbol{D}_i, \boldsymbol{C}_i, \boldsymbol{B}_i)\ dZ_i \\ &= \int_{Z_i} f(Y_i|X_i, Z_i, \boldsymbol{C}_i, \boldsymbol{B}_i)\ f(Z_i|X_i, \boldsymbol{D}_i, \boldsymbol{C}_i)\ dZ_i \end{aligned} \tag{1.2}$$

The first term under the integral is the TDM and the second is the MEM. While there would typically be either internal or external validation data, not both, the likelihood for n_m main study subjects, n_i internal validation subjects, and n_e external validation subjects would be:

$$\begin{aligned} L(\boldsymbol{\theta}) = &\left(\prod_{i=1}^{n_m} \int_{Z_i} f(Y_i|X_i, Z_i, \boldsymbol{C}_i, \boldsymbol{B}_i)\ f(Z_i|X_i, \boldsymbol{D}_i, \boldsymbol{C}_i)\ dZ_i\right) \\ &\left(\prod_{j=1}^{n_i} f(Y_j|X_j, Z_j, \boldsymbol{C}_j, \boldsymbol{B}_j)\ f(Z_j|X_j, \boldsymbol{D}_j, \boldsymbol{C}_j)\right) \\ &\left(\prod_{k=1}^{n_e} f(Z_k|X_k, \boldsymbol{D}_k, \boldsymbol{C}_k)\right) \end{aligned} \tag{1.3}$$

Several notable special cases are as follows. First, if the TDM and MEM are both normal-errors linear regressions:

$$\begin{aligned} &\text{TDM:}\ \ Y = \beta_0 + \beta_x X + \beta_z Z + \boldsymbol{\beta_c}^T \boldsymbol{C} + \boldsymbol{\beta_b}^T \boldsymbol{B} + \epsilon,\ \epsilon \sim N(0, \sigma_\epsilon^2) \\ &\text{MEM:}\ \ Z = \alpha_0 + \alpha_x X + \boldsymbol{\alpha_d}^T \boldsymbol{D} + \boldsymbol{\alpha_c}^T \boldsymbol{C} + \delta,\ \delta \sim N(0, \sigma_\delta^2) \end{aligned} \tag{1.4}$$

Then $Y|(X, \boldsymbol{D}, \boldsymbol{C}, \boldsymbol{B})$ is normal with:

$$\begin{aligned} E(Y|X, \boldsymbol{D}, \boldsymbol{C}, \boldsymbol{B}) &= \beta_0 + \beta_z\alpha_0 + (\beta_x + \beta_z\alpha_x)X + \beta_z\boldsymbol{\alpha_d}^T\boldsymbol{D} + (\beta_z\boldsymbol{\alpha_c}^T + \boldsymbol{\beta_c}^T)\boldsymbol{C} + \boldsymbol{\beta_b}^T\boldsymbol{B} \\ V(Y|X, \boldsymbol{D}, \boldsymbol{C}, \boldsymbol{B}) &= \sigma_\epsilon^2 + \beta_z^2\sigma_\delta^2 \end{aligned} \tag{1.5}$$

which means the likelihood for main study subjects in Eq. 1.2 has a closed form. Another special case is a logistic regression TDM and linear regression MEM:

$$\begin{aligned} &\text{TDM: } \text{logit}[P(Y=1)] = \beta_0 + \beta_x X + \beta_z Z + \boldsymbol{\beta_c}^T\boldsymbol{C} + \boldsymbol{\beta_b}^T\boldsymbol{B} \\ &\text{MEM: } Z = \alpha_0 + \alpha_x X + \boldsymbol{\alpha_d}^T\boldsymbol{D} + \boldsymbol{\alpha_c}^T\boldsymbol{C} + \delta,\ \delta \sim N(0, \sigma_\delta^2) \end{aligned} \tag{1.6}$$

The main study likelihood is a logistic-normal integral:

$$\int_{-\infty}^{\infty} p^y(1-p)^{1-y} \frac{1}{\sqrt{2\pi\sigma_\delta^2}} e^{-\frac{1}{2\sigma_\delta^2}(Z-\mu_z)^2}\ dZ \tag{1.7}$$

where $p = \text{logit}^{-1}(\beta_0 + \beta_x X + \beta_z Z + \boldsymbol{\beta_c}^T\boldsymbol{C} + \boldsymbol{\beta_b}^T\boldsymbol{B})$ and $\mu_z = \alpha_0 + \alpha_x X + \boldsymbol{\alpha_d}^T\boldsymbol{D} + \boldsymbol{\alpha_c}^T\boldsymbol{C}$. The integral represents $P(Y = y|X, \boldsymbol{D}, \boldsymbol{C}, \boldsymbol{B})$. We consider full ML with numerical integration as well as the probit approximation, which uses the fact that $\text{logit}^{-1}(x) \approx \Phi(\frac{x}{1.7})$ (where $\Phi(\cdot)$ is the standard normal CDF) to obtain the approximation:

$$P(Y = 1|X, \boldsymbol{D}, \boldsymbol{C}, \boldsymbol{B}) \approx \frac{e^t}{1+e^t} \tag{1.8}$$

where:

$$t = \frac{\beta_0 + \beta_x X + \beta_z\mu_z + \boldsymbol{\beta_c}^T\boldsymbol{C} + \boldsymbol{\beta_b}^T\boldsymbol{B}}{\sqrt{1 + \frac{\beta_z^2\sigma_\delta^2}{1.7^2}}} \tag{1.9}$$

This approach has been used by other authors, with favorable results (e.g. Lyles and Kupper (2013); Carroll *et al.* (2006)).

A third special case is a log-transformed linear model for the MEM:

$$\text{MEM: } \log(Z) = \alpha_0 + \alpha_x X + \boldsymbol{\beta_d^T D} + \boldsymbol{\beta_c^T C} + \delta,\ \delta \sim N(0, \sigma_\delta^2) \tag{1.10}$$

This MEM has two advantages: it accommodates skewness in the unmeasured confounder, and its nonlinearity permits identifiability of all TDM parameters even without any $\boldsymbol{D}$ variables (Carroll *et al.*, 2006).

Finally, we note that a binary Z is easy to accommodate in the 2-model ML setup, as the integral in Eq. 1.2 becomes a summation over $Z \in (0, 1)$ and numerical integration is not needed.

After obtaining $\hat{\boldsymbol{\theta}}$ via optimization procedures, a variance-covariance matrix, $\hat{V}(\hat{\boldsymbol{\theta}})$ can be estimated by taking the inverse of the estimated Hessian matrix of the log-likelihood function evaluated at $\hat{\boldsymbol{\theta}}$. Standard errors are taken as the square roots of the diagonal elements of $\hat{V}(\hat{\boldsymbol{\beta}})$.

3-model ML

An alternative likelihood approach is the three-density factorization:

$$\begin{aligned} L_i(\boldsymbol{\theta}) &= f(Y_i, X_i | \boldsymbol{D}_i, \boldsymbol{C}_i, \boldsymbol{B}_i) \\ &= \int_{Z_i} f(Y_i, X_i, Z_i | \boldsymbol{D}_i, \boldsymbol{C}_i, \boldsymbol{B}_i)\ dZ_i \\ &= \int_{Z_i} f(Y_i | X_i, Z_i, \boldsymbol{C}_i, \boldsymbol{B}_i)\ f(X_i | Z_i, \boldsymbol{C}_i, \boldsymbol{B}_i)\ f(Z_i | \boldsymbol{D}_i, \boldsymbol{C}_i)\ dZ_i \end{aligned} \tag{1.11}$$

In the second density of the final line, omitting $\boldsymbol{D}$ reflects an assumption that these variables do not inform X given Z. This is in line with the intended role of $\boldsymbol{D}$ in the 2-model ML scenario, in that it permits identifiability with external validation data and a normal linear model for Z given covariates (the third density here). With internal validation data or some other model for Z, $\boldsymbol{D}$ could be included in the model for X.

This approach requires specifying three rather than two models, but it aligns more squarely with a confounding-type DAG like Figure 1.1 (A), as opposed to the mediation-type DAG like Figure 1.1 (B). If investigators are conceptualizing Z (and other covariates) as affecting X, they might prefer specifying a model for X given covariates, and then an additional model for $Z|(\boldsymbol{D}, \boldsymbol{C})$.

1.2.4 Regression calibration

There are two procedures commonly referred to as regression calibration (RC). We term the two variants the "conditional expectation" view and the "algebraic" view (Rosner *et al.*, 1989). They give identical point estimates in certain scenarios, e.g. main study/external validation study designs with a linear regression MEM and a single D variable (Thurston *et al.*, 2003). We mostly focus on the conditional expectation version, although we use the algebraic version when the two are equivalent since it readily permits delta method-based standard errors.

For notational convenience, here we let X be absorbed into $\boldsymbol{C}$. Applied to unmeasured confounding, RC requires a model for the expectation of Z given $(\boldsymbol{D}, \boldsymbol{C})$. The usual approach is a linear regression:

$$E(Z) = \alpha_0 + \boldsymbol{\alpha}_{\boldsymbol{d}}^T \boldsymbol{D} + \boldsymbol{\alpha}_{\boldsymbol{c}}^T \boldsymbol{C} \tag{1.12}$$

If Z is skewed, a lognormal version might be preferred:

$$\log(Z) = \alpha_0 + \boldsymbol{\alpha}_{\boldsymbol{d}}^T \boldsymbol{D} + \boldsymbol{\alpha}_{\boldsymbol{c}}^T \boldsymbol{C} + \delta,\ \delta \sim N(0, \sigma_\delta^2) \tag{1.13}$$

This MEM has been used in the measurement error literature for handling multiplicative rather than additive errors (e.g. Lyles and Kupper (1997, 2013)).

For the conditional expectation RC, we use validation data to estimate the MEM param-

eters and then fit the TDM with $E(Z|\boldsymbol{D},\boldsymbol{C};\hat{\boldsymbol{\alpha}})$ in place of the missing Z's for main study subjects. For Eq. 1.13, Z given covariates is lognormal, so $E(Z|\boldsymbol{D},\boldsymbol{C}) = e^{\alpha_0+\boldsymbol{\alpha_d}^T\boldsymbol{D}+\boldsymbol{\alpha_c}^T\boldsymbol{C}+\frac{1}{2}\sigma_\delta^2}$. We use percentile bootstrap confidence intervals with 1,000 bootstrap samples (Efron and Tibshirani, 1986).

With no $\boldsymbol{D}$ and external validation data, TDM parameters are not identifiable under Eq. 1.12 because $E(Z|\boldsymbol{C};\hat{\boldsymbol{\alpha}})$ is a linear combination of $(1,\boldsymbol{C}^T)$, but are identifiable under Eq. 1.13. This coincides with the identifiability conditions for 2-model ML (see Section 1.2.3, pg. 9).

For algebraic RC with the Eq. 1.12 MEM and scalar D, main study data are used to fit the TDM with D in place of the missing Z to obtain $\hat{\boldsymbol{\beta}}^* = (\hat{\beta}_0^*, \hat{\beta}_z^*, \hat{\boldsymbol{\beta}}_{\boldsymbol{c}}^{*T}, \hat{\boldsymbol{\beta}}_{\boldsymbol{b}}^{*T})^T$, and validation data are used to fit the MEM and obtain $\hat{\boldsymbol{\alpha}}$. The correspondence between $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}$, which holds exactly for a linear TDM and approximately (under certain conditions) for a logistic TDM, is as follows (Kuha, 1994):

$$\begin{aligned}
\beta_0^* &= \beta_0 + \beta_z\alpha_0 \\
\beta_z^* &= \beta_z\alpha_d \\
\boldsymbol{\beta_c^*} &= \beta_z\boldsymbol{\alpha_c}^T + \boldsymbol{\beta_c}^T \\
\boldsymbol{\beta_b^*} &= \boldsymbol{\beta_b}
\end{aligned} \tag{1.14}$$

which can be written as a system of equations:

$$\begin{pmatrix} 1 & \alpha_0 & \mathbf{0}_{k_c}^T & \mathbf{0}_{k_b}^T \\ 0 & \alpha_d & \mathbf{0}_{k_c}^T & \mathbf{0}_{k_b}^T \\ \mathbf{0}_{k_c} & \boldsymbol{\alpha_c} & \boldsymbol{I}_{k_c} & \mathbf{0}_{k_c}\mathbf{0}_{k_b}^T \\ \mathbf{0}_{k_b} & \mathbf{0}_{k_b} & \mathbf{0}_{k_b}\mathbf{0}_{k_c}^T & \boldsymbol{I}_{k_b} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_z \\ \boldsymbol{\beta_c} \\ \boldsymbol{\beta_b} \end{pmatrix} = \begin{pmatrix} \beta_0^* \\ \beta_z^* \\ \boldsymbol{\beta_c^*} \\ \boldsymbol{\beta_b^*} \end{pmatrix} \qquad \text{or} \qquad \boldsymbol{A\beta} = \boldsymbol{\beta}^* \tag{1.15}$$

The $\boldsymbol{A}$ matrix is square and typically invertible, giving the RC estimator $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{A}}^{-1}\hat{\boldsymbol{\beta}}^* =$

$g(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\alpha}})$. A variance estimator can be obtained using the delta method and the fact that $V(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\alpha}})$ is a block-diagonal matrix with $V(\hat{\boldsymbol{\beta}}^*)$ and $V(\hat{\boldsymbol{\alpha}})$ on the diagonal.

For a logistic regression TDM, the RC approximation can be expected to perform well when either (1) $\beta_z^2 V(Z|\boldsymbol{D}, \boldsymbol{C})$ is small (e.g. < 0.5) or (2) the disease is rare and $Z|(\boldsymbol{D}, \boldsymbol{C})$ is normally distributed (Kuha, 1994).

When validation data are internal, a natural idea is to fit the disease model using the measured Z's for internal validation subjects and $E(Z|\boldsymbol{D}, \boldsymbol{C}; \hat{\boldsymbol{\alpha}})$ for main study participants. This may be inefficient because it treats Z the same whether it is observed or imputed. Spiegelman *et al.* (2001) suggest a more efficient method along the lines of Greenland (1988), but it requires that $\boldsymbol{\beta}$ is identifiable with external validation data. Briefly, one uses the internal validation data to obtain a set of estimates, $\hat{\boldsymbol{\beta}}^I$, then treats the validation data as external (ignoring the Y's) to obtain a main study/external validation set of RC estimates, $\hat{\boldsymbol{\beta}}^{RC}$. Using the fact that the two estimators are asymptotically uncorrelated, the final weighted estimates are given by $\hat{\beta}_j = w_j \hat{\beta}_j^{RC} + (1 - w_j)\hat{\beta}_j^I$ where $w_j = \frac{\hat{V}(\hat{\beta}_j^I)}{\hat{V}(\hat{\beta}_j^I) + \hat{V}(\hat{\beta}_j^{RC})}$.

1.2.5 Propensity score calibration

Stürmer *et al.* (2005) developed propensity score calibration (PSC) to correct for unmeasured confounding with a binary exposure and a main study/external validation study design. It is based on the idea of controlling for confounding via a propensity score, i.e. the estimated probability of exposure X given all covariates of interest $(\boldsymbol{Z}, \boldsymbol{C})$. The TDM is a model for Y vs. (X, G), where G is the propensity score defined as $P(X = 1|\boldsymbol{Z}, \boldsymbol{C})$.

Validation data are used to fit a model for the gold standard propensity score $G = P(X = 1|\boldsymbol{Z}, \boldsymbol{C})$, a model for an error-prone propensity score $G^* = P(X = 1|\boldsymbol{C})$, and a

linear model relating the two that can be viewed as a MEM:

$$E(G) = \lambda_0 + \lambda_x X + \lambda_{g*} G^* \tag{1.16}$$

Then, $\hat{G}^*$ is calculated for main study subjects, mapped to $\hat{G}$ using the fitted Eq. 1.16, and the TDM is fit with the predicted $\hat{G}$'s in place of the G's.

The approach is very similar to RC, with G^* playing the role of $\boldsymbol{D}$ in that it informs the TDM covariate G but is assumed to be independent of Y given (X, G). For intuition on this surrogacy assumption, Stürmer *et al.* (2007b) state that "surrogacy is violated when the direction of confounding in the exposure-disease association caused by the unobserved variable(s) differs from that of the confounding due to observed variables."

We note that the surrogacy assumption is necessary for identifiability with external validation data, because $E(G|X, G^*; \hat{\boldsymbol{\lambda}})$ is a linear combination of $(1, X, G^*)$, so the TDM design matrix would be less than full rank if G^* were included. But it could be relaxed with internal validation data by simply including G^* in the TDM. While Stürmer *et al.* (2007b) present methods to test the surrogacy assumption prior to performing PSC, we are not aware of prior literature the performance of PSC with G^* added to the TDM.

PSC is simple to apply and particularly appealing when there are several variables identified as potential unmeasured confounders. It is also usable in a scenario where RC and ML lack identifiability: when validation data are external and Z given covariates is a linear model with no $\boldsymbol{D}$. However, it requires a binary exposure, it does not produce regression coefficient estimates for other individual predictors (like propensity score adjustment in general), and its performance depends on a critical surrogacy assumption.

Stürmer *et al.* (2005) suggest using SAS macros for regression calibration developed by Rosner *et al.* (1989) to obtain point estimates and standard errors for PSC, although bootstrapping may be more appropriate to account for the fact that the propensity scores are estimated rather than known (Lunt *et al.*, 2012).

After observing generally poor performance of PSC in simulations with external validation data, we propose a modified version that incorporates the extra $\boldsymbol{D}$ variables that ML and RC rely on for identifiability. If $\boldsymbol{D}$ informs Z, it should also inform the gold standard propensity score G. Adding $\boldsymbol{D}$ to the MEM allows us to include G^* in the TDM, or in other words relax the usual surrogacy assumption of PSC. The modified MEM is:

$$E(G) = \lambda_0 + \lambda_x X + \boldsymbol{\lambda_d^T D} + \lambda_{g*} G^* \tag{1.17}$$

We propose an additional version of PSC for use when validation data are internal and there is no $\boldsymbol{D}$, aimed at avoiding regression calibration-related efficiency losses. We first relax the PSC surrogacy assumption by including G^* in the TDM. In fitting the TDM, main study subjects have $\hat{G}^*$'s but not G's, while validation subjects have G^*'s and G's. This is analogous to 2-model ML and RC with $\boldsymbol{Z} = G$, $\boldsymbol{C} = G^*$, and no $\boldsymbol{D}$ or $\boldsymbol{B}$. Fitting the TDM via 2-model ML rather than RC may help to upweight the observed G's relative to the imputed $\hat{G}$'s. This requires an additional assumption that the errors in 1.17 are normally distributed. We denote this estimator PSC (ML) to reflect the fact that it utilizes ML.

Implementation details

All analyses and simulations were performed using R version 3.5.0 (R Core Team, 2015). Our functions are included in the R package **meuc** (Van Domelen, 2018a), which can be installed from GitHub by loading **devtools** (Wickham *et al.*, 2018) and running:

```
install_github("vandomed/meuc")
```

For ML, likelihood functions were maximized using the R function *nlminb*, which uses a quasi-Newton Raphson algorithm. Starting values were set to 0.01 for regression coefficients and 1 or 10 for variance terms, whichever was closer to the true value; lower bounds for variance terms were set to 0.0001. The *hcubature* function in the package **cubature** v.

1.3-11 (Narasimhan and Johnson, 2017) was used for numerical integration; it performs h-adaptive integration as described by Berntsen *et al.* (1991) and Genz and Malik (1980). We obtained variance-covariance matrices by inverting the Hessian matrix at the MLEs, approximating Hessian matrices numerically via the *hessian* function in **pracma** v. 2.1.1 (Borchers, 2017).

Relevant ML functions in **meuc** include *ml_logistic_linear*, *ml_linear_linear*, *ml_logistic_logistic_linear*, and *ml_linear_logistic_linear*. The naming convention is the form of the TDM followed by the form of the secondary model(s), so for example *ml_logistic_linear* is for logistic regression TDM and linear regression MEM, as in Eq. 1.6.

RC is implemented in the functions *rc_cond_exp* and *rc_algebraic*; the latter is limited to scalar D and external validation data, but produces delta method standard errors, while the former is more general. PSC is implemented in *psc_cond_exp*, *psc_algebraic*, and *psc_algebraic_d*.

1.3 Results

1.3.1 Motivating example: low Vitamin D and fecundity

The Effects of Aspirin in Gestation and Reproduction (EAGeR) Study was a clinical trial aimed at determining whether daily low-dose aspirin lowers the rate of pregnancy loss. A total of 1,228 women age 18-40 who had one or two prior miscarriages and planned to become pregnant again participated in the study. Further details are provided by Schisterman *et al.* (2013).

The primary finding from EAGeR was that low-dose aspirin was not associated with live birth rate or pregnancy loss (Schisterman *et al.*, 2014). But data from the trial has been

used to explore a variety of other research questions as well: for example, whether leptin is associated with live birth rate (Zarek *et al.*, 2015), and whether C-reactive protein is associated with pregnancy loss (Mumford *et al.*, 2015).

We are interested in testing the hypothesis that low Vitamin D is associated with lower odds of becoming pregnant, over a time period covering six menstrual cycles, and we wish to control for maternal age, overweight status, and caloric intake. We suspect caloric intake is a confounder because consuming fewer calories should increase the likelihood of low Vitamin D, but also the likelihood of deficiencies in other nutrients that may be related to odds of becoming pregnant.

All of the variables except caloric intake are available in EAGeR. Data from a different study, BioCycle, can serve as an external validation sample. Caloric intake was measured in BioCycle as the average total energy intake from up to eight 24-hour dietary recalls (Gaskins *et al.*, 2009).

BioCycle was a longitudinal study on oxidative stress and hormone levels during the menstrual cycle. A total of 259 women age 18-44 participated. Because this validation data is external, we have to assume transportability, or that relationships among variables in BioCycle participants are the same as in EAGeR participants. It is impossible to directly test this assumption, but we can compare some basic characteristics across the two samples (Table 1.1). Note that this comparison and all subsequent analyses are for white women only, since budget constraints resulted in caloric intake (the suspected unmeasured confounder) only being measured for white women in BioCycle.

The obesity rate was higher in EAGeR than in BioCycle, and EAGeR subjects averaged a higher income. Additionally, the women in EAGeR all had one or two prior miscarriages and were trying to become pregnant; this was not true for women in BioCycle.

The logistic regression fit for low Vitamin D and odds of incident pregnancy, adjusted for age and overweight status, is shown in Table 1.2. Low Vitamin D was associated with

Table 1.1: Characteristics of women in EAGeR and BioCycle.

Variable	EAGeR (n = 995)	BioCycle (n = 89)	P
Age (years), M (SD)	28.9 (4.7)	28.5 (8.5)	0.62
BMI, n (%)			0.002
Normal weight	539 (54.2)	61 (68.5)	
Overweight	231 (23.2)	22 (24.7)	
Obese	225 (22.6)	6 (6.7)	
Education, n (%)			0.34
High school or less	111 (11.2)	7 (7.9)	
More than high school	884 (88.8)	82 (92.1)	
Income, n (%)			<0.001
Less than $40,000	301 (30.3)	33 (37.5)	
[$40,000, $75,000)	149 (15.0)	31 (35.2)	
$75,000 or more	544 (54.7)	24 (27.3)	
Smoking status, n (%)			0.32
Non-smoker	891 (90.0)	83 (93.3)	
Smoker	99 (10.0)	6 (6.7)	

Table 1.2: Logistic regression estimates for odds of pregnancy in EAGeR (n = 995).

Variable	Beta (SE)	OR (95% CI)	P
Intercept	2.35 (0.46)	–	<0.001
Vitamin D < 30 ng/mL	-0.34 (0.15)	0.71 (0.53, 0.95)	0.02
Age (years)	-0.04 (0.02)	0.97 (0.94, 0.99)	0.02
Overweight	-0.32 (0.15)	0.73 (0.54, 0.97)	0.03

an estimated 39% lower odds of becoming pregnant (p = 0.02).

Next, we incorporate data from BioCycle to estimate the odds ratio for low Vitamin D after additionally adjusting for caloric intake. Applying PSC as proposed by Sturmer et al. is straightforward, while 2-model ML, 3-model ML, and RC require specification of measurement error models.

Because BioCycle is an external validation dataset, if the MEM is a linear regression, 2-model ML, 3-model ML, and RC require having at least one additional predictor (D) in the MEM but not in the TDM. A log-transformed linear regression MEM would alleviate this issue, but that was not supported by the data. In normal linear regressions for caloric intake vs. low Vitamin D, age, and overweight status (the MEM for 2-model ML and RC), AIC was 507.6 with the log transformation and 506.2 without. The methods could still be applied with log-transformed MEM's and no D, but the resulting estimates may be unreliable due to near lack of identifiability. The $\widehat{\log\text{-OR}}$ (95% CI) for low Vitamin D were as follows: 2-model ML (full), 0.74 (0.31, 1.73); 2-model ML (approximate), 0.76 (0.24, 2.43); 3-model ML, 0.74 (0.34, 1.61); and RC, 64.0 (0.00, Inf). Clearly the RC estimate is not useful here.

We identified height as a potentially useful additional variable (D) for the non-log-transformed MEM's, on the basis that height should inform caloric intake for a given BMI (which overweight status is based on) and can reasonably be assumed unrelated to odds of becoming pregnant. The fitted MEM for 2-model ML and RC is summarized in Table 1.3.

Height was indeed associated with caloric intake, while low Vitamin D, age, and over-

Table 1.3: Linear regression estimates for caloric intake in BioCycle (n = 89, $R^2 = 0.09$).

Variable	Beta (SE)	P
Intercept	-14.19 (12.37)	0.25
Height (cm)	0.18 (0.08)	0.02
Vitamin D < 30 ng/mL	0.70 (0.85)	0.41
Age (years)	0.06 (0.05)	0.26
Overweight	0.81 (0.92)	0.38

weight status were not; these variables were also not informative of caloric intake in a model fit without height (all $p > 0.1$; not shown). The lack of an association between low Vitamin D and caloric intake suggests caloric intake is likely not an important confounder.

The assumed models for the various corrective methods are as follows, with Y = incident pregnancy, X = low Vitamin D, Z = caloric intake, C_1 = age, C_2 = overweight status, and D = height. For 2-model ML:

$$\begin{aligned} \text{TDM:}\ & \text{logit}[P(Y=1)] = \beta_0 + \beta_x X + \beta_z Z + \beta_{c_1} C_1 + \beta_{c_2} C_2 \\ \text{MEM:}\ & Z = \alpha_0 + \alpha_x X + \alpha_d D + \alpha_{c_1} C_1 + \alpha_{c_2} C_2 + \delta,\ \delta \sim N(0, \sigma_\delta^2) \end{aligned} \tag{1.18}$$

For 3-model ML (α's are implied to be different from Eq. 1.18).

$$\begin{aligned} \text{TDM:}\ & \text{logit}[P(Y=1)] = \beta_0 + \beta_x X + \beta_z Z + \beta_{c_1} C_1 + \beta_{c_2} C_2 \\ & \text{logit}[P(X=1)] = \gamma_0 + \gamma_z Z + \gamma_{c_1} C_1 + \gamma_{c_2} C_2 \\ \text{MEM:}\ & Z = \alpha_0 + \alpha_d D + \alpha_{c_1} C_1 + \alpha_{c_2} C_2 + \delta,\ \delta \sim N(0, \sigma_\delta^2) \end{aligned} \tag{1.19}$$

For RC:

$$\begin{aligned} \text{TDM:}\ & \text{logit}[P(Y=1)] = \beta_0 + \beta_x X + \beta_z Z + \beta_{c_1} C_1 + \beta_{c_2} C_2 \\ \text{MEM:}\ & E(Z) = \alpha_0 + \alpha_x X + \alpha_d D + \alpha_{c_1} C_1 + \alpha_{c_2} C_2 \end{aligned} \tag{1.20}$$

For standard PSC (G and G^* are gold standard and error-prone propensity scores, respec-

Table 1.4: Logistic regression estimates for odds of pregnancy using data from EAGeR and BioCycle. Models are also adjusted for age and overweight status.

	OR (95% CI)	
	Vitamin D < 30 ng/mL	Calories (Δ 100 kcal/day)
Naive		
Covariate adjustment	0.71 (0.53, 0.95)	-
Propensity score adjustment	0.71 (0.53, 0.95)	-
Corrected		
2-model ML (full)	0.71 (0.52, 0.97)	1.00 (0.89, 1.12)
2-model ML (approximate)	0.71 (0.52, 0.97)	1.00 (0.88, 1.12)
3-model ML	0.71 (0.52, 0.98)	1.00 (0.89, 1.12)
Regression calibration	0.71 (0.52, 0.97)	1.00 (0.88, 1.12)
Propensity score cal.	0.73 (0.47, 1.07)	-
Propensity score cal. with D	0.72 (0.47, 1.03)	-

tively):

$$\begin{aligned} \text{TDM:}\ \ & \text{logit}[P(Y=1)] = \beta_0 + \beta_x X + \beta_g G \\ \text{MEM:}\ \ & E(G) = \lambda_0 + \lambda_x X + \lambda_{g^*} G^* \end{aligned} \tag{1.21}$$

And for PSC with surrogacy relaxed by adding height to the MEM:

$$\begin{aligned} \text{TDM:}\ \ & \text{logit}[P(Y=1)] = \beta_0 + \beta_x X + \beta_g G + \beta_{g^*} G^* \\ \text{MEM:}\ \ & E(G) = \lambda_0 + \lambda_x X + \lambda_d D + \lambda_{g^*} G^* \end{aligned} \tag{1.22}$$

Table 1.4 shows the estimated odds ratios for low Vitamin D and for caloric intake for various corrective methods using data from both EAGeR and BioCycle. Estimates are very similar for all methods, likely stemming from the lack of association between low Vitamin D and the suspected unmeasured confounder, caloric intake. But the corrective methods permitted checking for confounding due to caloric intake, while only slightly reducing the precision of the estimated exposure effect.

1.3.2 Simulations

Mimicking EAGeR/BioCycle (external validation)

The first set of simulations is intended to assess validity and efficiency of the various methods under conditions mimicking EAGeR/BioCycle, but with slightly modified parameters to induce confounding due to caloric intake. For each trial, we generate data for 995 main study subjects and 89 external validation subjects as follows: age, $C_1 \in (18, 19, ..., 44)$ with sampling probabilities equal to the sample proportions in EAGeR/BioCycle; overweight status, $C_2 \sim$ Bernoulli(0.45); low Vitamin D, $X \sim$ Bernoulli(logit$^{-1}(0.14 + 0.41C_2)$); and height, $D \sim N(166.3, 6.7^2)$. The MEM for $Z =$ caloric intake and TDM for $Y =$ incident pregnancy were as follows:

$$\begin{aligned} &\text{MEM: } Z = -14.18 - 1.00X + 0.18D + 0.06C_1 + 0.81C_2 + \delta,\ \delta \sim N(0, 7.5) \\ &\text{TDM: } \text{logit}[P(Y = 1)] = -0.32 - 0.34X + 0.15Z - 0.03C_1 - 0.32C_2 \end{aligned} \tag{1.23}$$

Parameter values were set to the estimated values from 2-model ML, with the following exceptions: the TDM coefficient for caloric intake, β_z, was set to 0.15 rather than 0.00; the MEM coefficient for low Vitamin D, α_x, -1 rather than 0.70; the MEM residual error variance, σ^2_δ, 7.5 rather than 14.98 to produce a plausible range of caloric intake values; and the TDM intercept, β_0, -0.32 rather than 2.40 to maintain a pregnancy incidence of 0.73 as it was in EAGeR. Note that the outcome is not rare, so RC performance depends on $\beta^2_z\sigma^2_\delta$ being small. That quantity is $(0.15^2)(7.5) = 0.17$, which is below the cut-off of 0.5 suggested by Kuha (1994), so we expect good performance for RC.

With height as a surrogate, we used the algebraic version of RC with delta method standard errors. We took a similar approach for the standard and modified PSC estimators, where $G^* =$ the error-prone propensity score and $D =$ height play the role of surrogate,

Table 1.5: Simulation results for estimation of adjusted log-OR for low Vitamin D and incident pregnancy with external validation data (1,000 trials, 1 with $|\hat{\beta}_x| > 1.5$ for 2-model ML estimators excluded and 12 with $|\hat{\beta}_x| > 1.5$ for 3-model ML excluded; true log-OR = -0.34).

	Mean bias	SD	Mean SE	MSE	Coverage
Unobservable truth					
Covariate adjustment	-0.013	0.155	0.155	0.024	0.955
Propensity score adjustment	-0.009	0.153	0.154	0.023	0.958
Naive					
Covariate adjustment	-0.144	0.151	0.151	0.043	0.842
Propensity score adjustment	-0.142	0.150	0.151	0.043	0.844
Corrected					
2-model ML (full)	-0.012	0.202	0.209	0.041	0.963
2-model ML (approximate)	-0.011	0.200	0.207	0.040	0.964
3-model ML	-0.101	0.204	0.213	0.052	0.909
Regression calibration	0.000	0.193	0.201	0.037	0.971
Propensity score cal.	-0.108	0.308	0.237	0.106	0.874
Propensity score cal. with D	-0.002	0.196	0.203	0.038	0.966

respectively. Simulation results for 1,000 trials are summarized in Table 1.5.

The naive exposure estimates were biased away from the null for both covariate adjustment and propensity score adjustment. The 2-model ML estimators performed well, which is expected given that they correspond to a correctly specified likelihood; the two versions produced very similar estimates ($r = 0.9997$). 3-model ML exhibited considerable bias and was somewhat unstable, perhaps due to difficulty estimating the $X|(Z, \boldsymbol{C})$ logistic regression parameters with only 89 validation study subjects. RC performed very well, with no bias and better efficiency than 2-model ML. The standard PSC without incorporating height exhibited bias away from the null nearly as severe as the naive estimates, while PSC modified with height in the MEM was unbiased and nearly as efficient as RC.

Standard deviations were only moderately larger for the corrective methods than for

the unobservable truth estimators, suggesting that 2-model ML, RC, and PSC with D can correct for bias associated with the unmeasured confounder in this scenario with only a modest loss of precision relative to actually measuring Z in the main study. Estimation of β_z, on the other hand, was much less precise for the corrective methods (excluding 1 trial with 2-model ML $\hat{\beta}_z$'s > 1.5: SD = 0.025 for unobservable truth covariate adjustment, 0.098 for 2-model ML (full), 0.094 for 2-model ML (approximate), and 0.077 for RC).

To assess performance of the methods for larger unmeasured confounding effects, we ran additional simulations with β_z ranging from -0.5 to 0.5, with β_0 also adjusted to maintain a pregnancy incidence of 0.73. Results are summarized in Figure 1.2. The 2-model ML (approximate) estimator of β_x performed well over a range of β_z values, while RC exhibited upward bias for large $|\beta_z|$. The standard PSC estimator was badly biased, while PSC with D performed reasonably well, but broke down similarly to RC for large $|\beta_z|$.

While internal validation data with surrogates is not a central focus, we ran additional simulations to quantify the efficiency gain associated with having internal rather than external validation data in this particular scenario. In 1,000 trials with internal validation data, the 2-model ML (approximate) estimator had fewer extreme estimates (0 vs. 1 trial with $|\hat{\beta}_x| > 1.5$) and was more precise (SD = 0.186 vs. 0.200) than with external validation data.

Mimicking EAGeR/BioCycle (internal validation)

Next we consider a similar scenario with internal rather than external validation data. In addition to not having to assume transportability (see Section 1.2.2, pg. 8), a key advantage of internal validation data is not requiring $\boldsymbol{D}$ for identifiability, so we also omit D = height from these simulations. The MEM and TDM are as follows:

$$\begin{aligned} &\text{MEM: } Z = 14.51 - 1.00X + 0.07C_1 + 0.69C_2 + \delta,\ \delta \sim N(0, 7.5) \\ &\text{TDM: } \text{logit}[P(Y=1)] = -0.18 - 0.34X + 0.15Z - 0.03C_1 - 0.32C_2 \end{aligned} \tag{1.24}$$

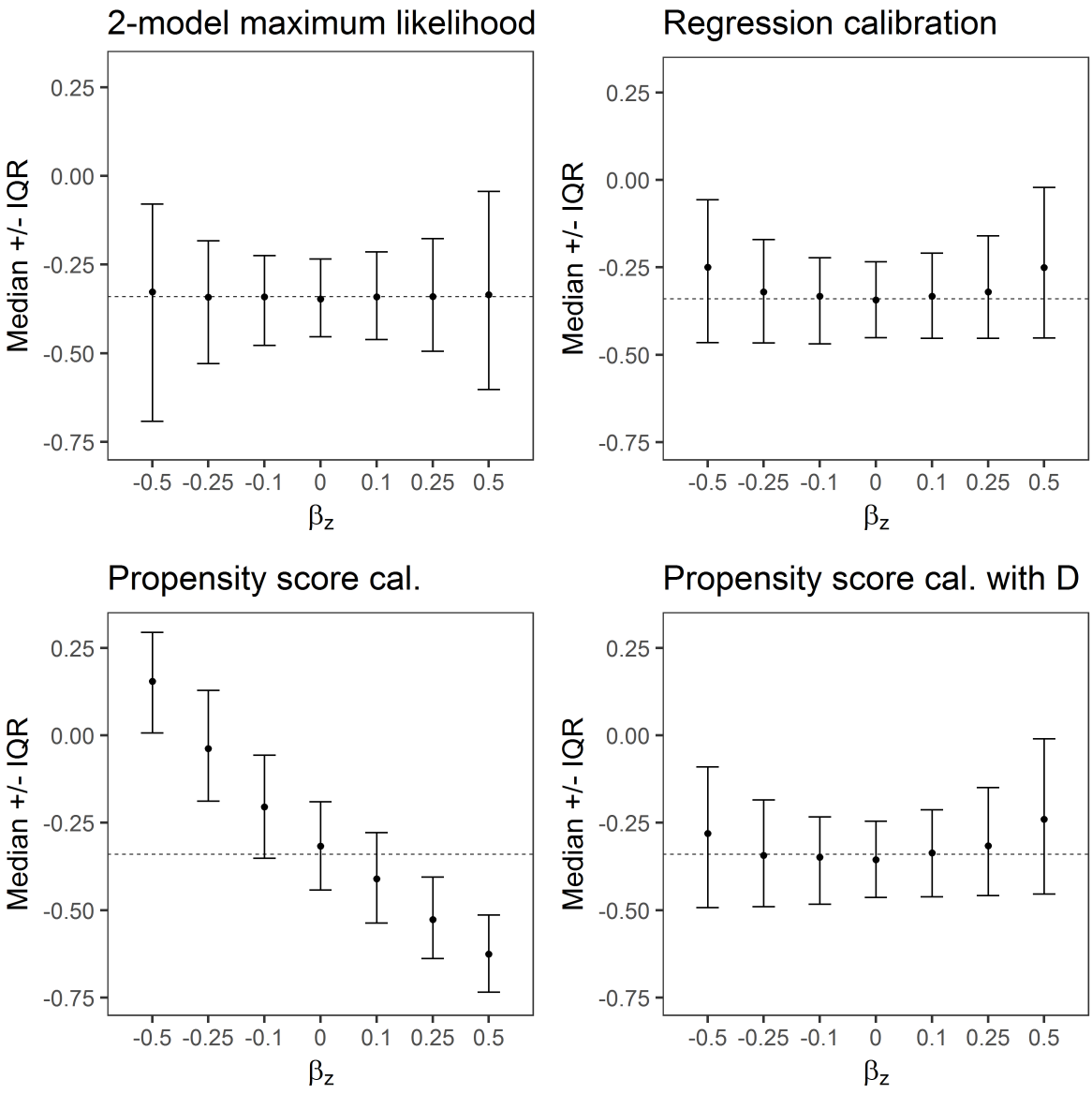

Figure 1.2: Performance of corrective methods as β_z varies (2,000 trials each).

Table 1.6: Simulation results for estimation of adjusted log-OR for low Vitamin D and incident pregnancy with internal validation data (1,000 trials, true log-OR = -0.34).

	Mean bias	SD	Mean SE	MSE	Coverage
Unobservable truth					
Covariate adjustment	-0.002	0.158	0.155	0.025	0.957
Propensity score adjustment	0.000	0.156	0.154	0.024	0.958
Naive					
Covariate adjustment	-0.139	0.144	0.145	0.040	0.845
Propensity score adjustment	-0.137	0.152	0.151	0.042	0.850
Internal validation subset					
Covariate adjustment	-0.047	0.592	0.561	0.352	0.952
Propensity score adjustment	-0.034	0.563	0.548	0.318	0.955
Corrected					
2-model ML (full)	0.006	0.204	0.206	0.042	0.966
2-model ML (approximate)	0.007	0.204	0.206	0.042	0.967
3-model ML	0.005	0.203	0.206	0.041	0.966
Regression calibration	0.014	0.198	0.216	0.039	0.959
Propensity score cal.	-0.083	0.174	0.183	0.037	0.952
Propensity score cal. w/o surr.	0.008	0.200	0.219	0.040	0.959

For RC, we used the observed Z's for internal validation subjects and $\hat{Z}$'s for main study subjects when fitting the TDM, and used percentile bootstrap confidence intervals. We took a similar approach for PSC, using the gold standard propensity scores for internal validation subjects and the predicted ones for main study subjects, along with bootstrap confidence intervals. We also included a PSC estimator with the surrogacy assumption relaxed by including both the gold standard and error-prone propensity scores in the TDM. Results are summarized in Table 1.6.

The internal validation subset estimators had high variability and exhibited bias away from the null, likely stemming from the small sample size ($n_i = 89$) and high outcome incidence (0.73). For the corrective methods, the 2-model ML estimators performed well,

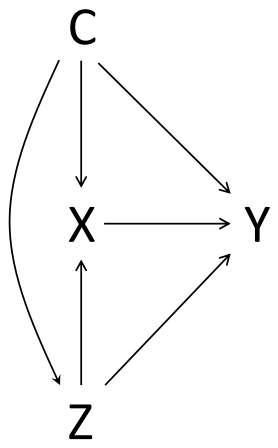

Figure 1.3: DAG for linear regression simulations.

again as expected since they are based on a correctly specified likelihood. In contrast to the previous scenario (Table 1.5), 3-model ML performed very similar to 2-model ML here. RC had slight upward bias but good coverage, and was slightly more efficient than the ML estimators. The standard PSC exhibited bias away from the null, but not as severe as in the prior scenario; it reduced bias relative to the naive estimators considerably. PSC with surrogacy relaxed performed similarly to ML and RC in terms of bias, efficiency, and CI coverage.

Linear regression with 3-model setup

Next, we consider a linear regression rather than logistic regression TDM, with a 3-model data generating process as shown in the Figure 1.3 DAG. This scenario corresponds to 3-model ML described in Section 1.2.3 with no $\boldsymbol{B}$ or $\boldsymbol{D}$ variables.

The effect of X on Y is confounded by Z and C, and control for (Z, C) is sufficient to remove confounding. With Z missing for main study subjects, the 2-model ML and RC approaches of the previous section could be used as a working model for the data, while a 3-model ML would be more in line with the relationships among variables illustrated in Figure 1.3.

Table 1.7: Simulation results for linear regression scenario with external validation data (1,000 trials, true $\beta_x = 0.5$).

	Median bias	IQR	Coverage
Unobservable truth			
Covariate adjustment	0.000	0.093	0.960
Propensity score adjustment	0.000	0.092	0.963
Naive			
Covariate adjustment	0.209	0.097	0.188
Propensity score adjustment	0.210	0.098	0.187
Corrected			
Propensity score calibration	-0.104	0.513	0.873

We use the following data generating process:

$$\begin{aligned} C &\sim N(0,1) \\ Z &= 0.25C + \delta,\ \delta \sim N(0,1) \\ \text{logit}[P(X=1)] &= \log(1.75)Z + \log(1.25)C \\ Y &= 0.4X + 0.5Z + 0.3C + \epsilon,\ \epsilon \sim N(0, 0.5) \end{aligned} \tag{1.25}$$

With external validation data, PSC is the only corrective method with identifiability; 2-model ML, 3-model ML, and RC would require at least one D variable in their respective MEM's. Simulation results for $n_m = 500$ and $n_e = 100$ are shown in Table 1.7. PSC performed poorly, overcorrecting for the confounding effect of Z and giving very imprecise estimates.

Results for internal validation data with $n_m = 500$ and $n_i = 100$ are summarized in Table 1.8. For PSC, the first two estimators in Table 1.8 are the same as those in Table 1.6, while the third is aimed at avoiding RC-related efficiency losses relative to ML with internal validation data (see Section 1.2.5, pg. 14).

The correctly specified likelihood was 3-model ML, but 3-model ML and 2-model ML

Table 1.8: Simulation results for linear regression scenario with internal validation data (1,000 trials, 3 with non-positive definite variance-covariance matrix for PSC (ML) excluded; true $\beta_x = 0.5$)

	Mean bias	SD	Mean SE	MSE	Coverage
Unobservable truth					
Covariate adjustment	-0.001	0.067	0.067	0.005	0.946
Propensity score adjustment	-0.001	0.068	0.068	0.005	0.950
Naive					
Covariate adjustment	0.207	0.067	0.067	0.047	0.122
Propensity score adjustment	0.207	0.074	0.073	0.048	0.196
Internal subset					
Covariate adjustment	0.002	0.141	0.150	0.020	0.963
Propensity score adjustment	0.002	0.141	0.157	0.020	0.971
Corrected					
3-model ML	-0.005	0.092	0.093	0.008	0.950
2-model ML	-0.006	0.092	0.093	0.008	0.959
Regression calibration	-0.001	0.102	0.103	0.010	0.947
Propensity score cal.	-0.041	0.168	0.174	0.030	0.985
Propensity score cal. w/o surr.	-0.001	0.103	0.104	0.011	0.944
Propensity score cal. (ML)	-0.006	0.092	0.093	0.009	0.956

performed very similarly ($r = 0.9983$) and had almost identical performance metrics in Table 1.8. The ML estimators had slight downward bias but were somewhat more efficient than RC; all three had valid standard errors and good CI coverage.

The standard PSC procedure reduced bias considerably relative to the naive estimators. However, its performance was very poor considering that it was more biased and less efficient than the estimators using just the internal validation subset. PSC with the surrogacy assumption relaxed performed well, and was very similar to RC. Implementing the PSC procedure via ML rather than RC increased efficiency, and resulted in similar performance to 3-model ML and 2-model ML.

The 3 trials in which PSC (ML) produced a non-positive definite variance-covariance matrix seemed to correspond to trials in which Z by chance did not inform X, resulting in almost perfect agreement between G^* and G for internal validation study subjects.

A final note, RC estimates of β_z were identical to β_z estimates using just the internal validation sample, while ML estimates were much more efficient (SD = 0.064 for 3-model ML and 2-model ML, SD = 0.075 for internal validation subset).

1.4 Discussion

As analysts increasingly leverage data from existing sources to explore new research questions, the problem of unmeasured confounding is becoming more common. When a potentially important confounder is missing, we propose seeking validation data and using methods from the measurement error literature to restore validity. In our simulations, RC and ML typically outperformed PSC, which was developed for unmeasured confounding and also relies on validation data.

The performance of PSC in main study/external validation scenarios, which is what it was developed for and where it offers unique identifiability, was disappointing. It reduced

bias relative to naive estimates ignoring the missing Z in all scenarios, but sometimes only modestly, and it sometimes overcorrected. This is consistent with previous simulation studies involving the method (Stürmer *et al.*, 2005, 2007b,a). In a relatively simple linear regression scenario with internal validation data, it was biased and less efficient than simply fitting the TDM with the internal validation data (see Table 1.8).

However, we proposed several modifications to PSC that improved its performance in various scenarios. Each relies on relaxing the critical surrogacy assumption of PSC, i.e. including the error-prone propensity score in the TDM in addition to the gold standard propensity score. First, if there is a variable that can reasonably be assumed to inform Z but not Y, as height was assumed to inform caloric intake but not odds of pregnancy in the motivating example, then including that variable in the $G|(X, G^*)$ measurement error model allows G^* to be added to the TDM. This drastically improved performance of PSC in simulations, but also nullified a key advantage of PSC relative to ML and RC in terms of identifiability with external validation data.

A second modification was simply including G^* in the TDM when validation data were internal. This scenario was examined by Stürmer *et al.* (2007b), who suggested two approaches for testing surrogacy: a likelihood ratio test for the regression coefficient for G^* in the TDM, and the proportion of the variance in Y explained by (G, G^*) that is due to G (close to 1 favors surrogacy). We did not assess either approach, partly because our motivating example had external validation data. But considering that the standard PSC procedure was biased in every external validation scenario we considered, it would seem advisable to relax surrogacy whenever possible. Conversely, including G^* in the TDM when it is not necessary could inflate standard errors for the estimated exposure effect, especially since G and G^* may be highly correlated.

A third modification was fitting the TDM via ML rather than RC, again for the scenario where validation data are internal and surrogacy is relaxed. In linear regression simulations,

this prevented RC-related efficiency losses and led to a PSC estimator equally efficient as ML. We suspect the same approach could prevent RC-related bias in logistic regression when the RC conditions break down (e.g. Figure 1.2), although we did not confirm this. Using ML for PSC estimation also permits delta method standard errors rather than bootstrapping. This approach is not entirely justified since it treats the propensity scores as fixed, but it worked well in our scenario (e.g. good CI coverage in Table 1.8).

In general, we favor ML over RC and PSC for adjusting for an unmeasured confounder with validation data. First, note that PSC and RC are both limited in that they require certain types of variables: a binary exposure for PSC, and a continuous (potentially skewed) unmeasured confounder for RC. ML is more flexible. Second, RC is an approximate method for TDMs other than linear regression, and may break down under certain conditions. These are well-known for logistic regression (Kuha, 1994) but less so for other models. Theoretical properties of PSC are not well-defined. Stürmer *et al.* (2007b) state that "surrogacy is a sufficient but not always necessary condition for PSC to be valid," while also reporting that it "had a tendency to overadjust even in scenarios where surrogacy was met.'" Third, while RC or a special version of it (e.g. Spiegelman *et al.* (2001)) is fully efficient relative to ML in certain scenarios, we are not aware of an efficient version of RC for internal validation data with no surrogates. This situation is more likely to occur in unmeasured confounding than in measurement error scenarios, where there is usually an imprecise version of the missing covariate.

Another disadvantage of PSC is that it only produces an effect estimate for one designated exposure. In some cases, investigators may wish to interpret adjusted associations for all of the covariates in the disease model; there may not even be a particular covariate that is viewed as the exposure of interest.

There are some computational aspects of the PSC procedure to examine more closely in future work. For example, the original procedure proposed by Stürmer *et al.* (2005) uses

only the validation data to fit the error-prone propensity score model. But $(X, \boldsymbol{C})$ are also observed for main study subjects, leading to several alternatives. One could include main study data in fitting the error-prone propensity score model, or perhaps fit the model twice, first with validation data to estimate MEM parameters, and again with main study data to obtain propensity scores for those subjects. Preliminary simulations suggest that fitting the model separately leads to more precise $\hat{\beta}_x$ estimates when the validation dataset is smaller than the main dataset, while the original approach is more efficient when the validation dataset is larger.

Identifiability is an important issue for all three of the corrective methods we examined. In general, ML and RC have identifiability when any of three conditions are met: (1) validation data are internal, (2) there are one or more surrogate variables $\boldsymbol{D}$ in the model for Z (the "MEM" for 2-model ML and RC) but not in the TDM; or (3) the model for Z is *not* a linear regression. For (2), we note that the surrogate could be a higher-order term involving the TDM covariates rather than an entirely distinct variable. For example, in the EAGeR/BioCycle analysis, if there was a significant age-by-overweight status interaction on caloric intake, that interaction term could have been used for D rather than height. For (3), skewness in Z could be exploited to permit identifiability, e.g. by using a log-transformed MEM as used in the measurement error literature for handling multiplicative errors (Lyles and Kupper, 1997). Caloric intake was not skewed in BioCycle, so we did not pursue this approach, but it may be worth exploring in future work.

Identifiability conditions are similar for PSC if relaxing its surrogacy assumption is viewed as a requirement for valid estimation. Surrogacy can be relaxed if validation data are internal or there are one or more variables in the MEM for $G|(X, G^*)$ but omitted from the TDM.

In handling covariate measurement error, there are numerous compelling reasons to prefer internal rather than external validation data: for example, no need to assume transportability, and more efficient estimation for the same validation study sample size (Carroll *et al.*,

2006). For unmeasured confounding, identifiability without any surrogate variables is an additional key advantage. In secondary data analysis, the validation dataset on hand will simply either be internal or external, while in original investigations one might have control over the nature of the validation data. If a potential confounder is realized after data collection is partly or even mostly complete, it would be extremely useful to simply start measuring the variable for the remaining subjects, as opposed to designing a separate validation study or utilizing existing validation data from another study. This will typically require IRB approval, but perhaps only a minor modification that can be approved quickly, depending on the nature of the variable and how it is measured.

In other cases, the optimal approach may be less clear. If an existing dataset is missing an important covariate, one might consider designing a validation study to combine with the primary dataset. Once the decision is made to pursue a main study/validation study correction, it is certainly preferred to collect outcome data in the validation study if feasible. On the other hand, while the validation dataset would be internal in the sense that it contains outcome data, it would be external in the sense that it was collected separately, likely in a different geographical location. So validity would still rest on a transportability assumption. In certain circumstances, particularly when there are concerns about transportability, it may be a better use of resources to design a new study measuring all relevant variables rather than a validation study to complement an existing dataset. Utilizing the existing dataset would enhance precision of parameter estimates if the assumption is met, but may compromise validity if it is not (Carroll *et al.*, 2006).

In summary, maximum likelihood and regression calibration approaches developed for covariate measurement error are well suited for the unmeasured confounding problem. In our simulations, ML and RC generally outperformed the original PSC procedure proposed by Stürmer *et al.* (2005). We identified several ways to modify PSC to improve its performance, e.g. by relaxing its surrogacy assumption and by fitting the disease model of interest via

ML rather than RC to avoid efficiency losses. All of these methods are implemented in the R package **meuc** (Van Domelen, 2018a).

Chapter 2: Estimating the covariate-adjusted log-odds ratio for a continuous exposure measured in pools and subject to errors

2.1 Introduction

In pooling studies, a biomarker of interest is measured in combined biospecimen samples from multiple subjects rather than for each individual (Dorfman, 1943; Chen *et al.*, 2009). Pooling designs may be best known for their use in determining individual-level disease status with fewer assays, e.g. screening donated blood for hepatitis B virus (Stramer *et al.*, 2013). But our focus is on the use of pooling for measuring a continuous biomarker and estimating parameters in an individual-level regression model of interest.

There are numerous reasons to consider utilizing a pooling study design. The assay of interest may require volumes of sample greater than available for individual subjects in a given study (Weinberg and Umbach, 1999). In a regression setting, if pools are comprised of samples from subjects similar on relevant characteristics, a pooling design requiring many fewer assays may offer only slightly less power than a traditional design where each subject's

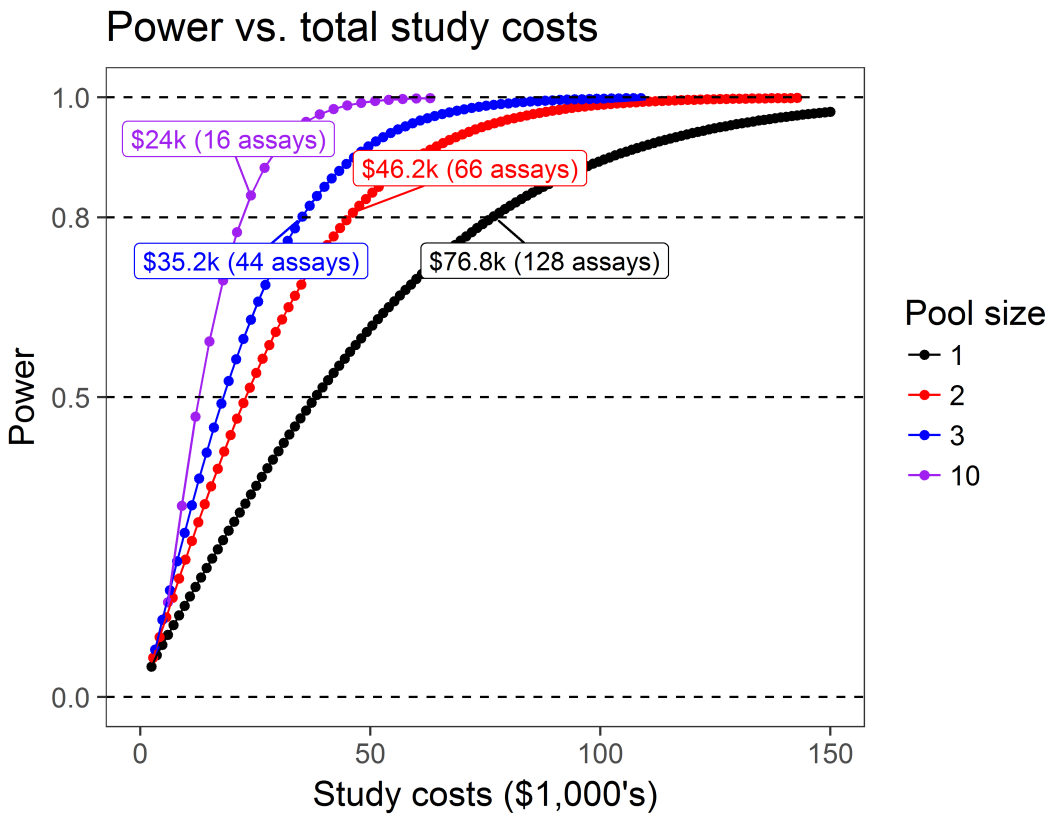

Figure 2.4: Power vs. total study costs for hypothetical two-sample t-test scenario.

biomarker level is measured (Lyles *et al.*, 2016; Mitchell *et al.*, 2014b). A pooling design may be able to achieve much higher power for the same budget, or the same power for a much lower total cost.

Figure 2.4 illustrates pooling-related efficiency gains for a two-sample t-test, where the true mean difference is 0.5, the variance of biomarker levels is 1, each assay costs \$500, and other costs per subject total \$100. Achieving 80% power with individual assays requires 64 subjects per group, at a total cost of \$76,800; with pools of size 10, 80% power requires 80 subjects per group, with 16 total assays and a cost of \$24,000. For the same cost as the traditional design with 80% power, pooling designs with 2, 3, and 10 members per pool achieve powers of 95.4%, 98.7%, and >99.9%, respectively.

Pooling can be used for highly efficient estimation of the log-odds ratio relating a binary

outcome to a continuous exposure, provided pools can be formed to be homogeneous with respect to case status. Weinberg and Umbach (1999) provide a logistic regression model that can be used to estimate the log-odds ratios of interest with poolwise sums rather than individual measurements for the exposure and covariates. As they note, however, fitting this model without accounting for errors in the biomarker measurements can lead to inconsistent parameter estimation (Carroll *et al.*, 2006; Fuller, 1987).

Schisterman *et al.* (2010) describe two types of errors that can contaminate pooled biomarker measurements and potentially induce bias: measurement error due to assay-related imprecision, and processing or "pooling" error due to the physical process of combining samples. Processing error can be caused by imperfect lab conditions, unintentionally unequal specimen volumes, and cross-reactions between components of blood from different people (Schisterman *et al.*, 2010; Weinberg and Umbach, 1999). Schisterman *et al.* (2010) focused on estimating parameters of a biomarker distribution via a hybrid study design with several different pool sizes including 1 (i.e. single measurements), using the fact that different pool sizes are subject to different combinations of error types.

Lyles *et al.* (2015) adopted the Schisterman *et al.* (2010) framework to estimate a covariate-adjusted log-odds ratio with poolwise data while correcting for errors. They used a discriminant function approach, targeting the log-odds ratio of interest via a normal-errors regression of the biomarker on case status and covariates rather than a logistic regression. The individual-level discriminant function model implies a similar poolwise model, into which additive normal errors can be incorporated, resulting in a closed-form likelihood for the pooled observations. This approach is computationally simple and does not require homogeneous pools, but it only produces a log-odds ratio estimate for the pooled biomarker, not for covariates. The likelihood methods of Liu *et al.* (2017) are similar in that they model the pooled biomarker as the dependent variable, but their focus is on outcome measurement error, as opposed to correcting an adjusted log-odds ratio estimate for covariate errors.

In this chapter, we follow the framework of Schisterman *et al.* (2010) and Lyles *et al.* (2015) to extend the Weinberg and Umbach (1999) logistic regression model to accommodate errors in the pooled exposure. We consider a hybrid study design that includes several different pool sizes, typically including some singles. We pursue likelihood-based inference assuming processing and measurement errors are independent and normally distributed with 0 means and variances that do not depend on the pool size. We assume processing error affects pools of size ≥ 2 only, while measurement errors affect all observations.

While all parameters are identifiable with a design that includes at least three different pool sizes including 1, we demonstrate that numerical stability and precision can be improved by including a small number of replicates in the study design. By replicates, we mean specifically that two assay measurements are obtained for some single-specimen pools. We apply our methods to explore the relationship between levels of a serum cytokine during pregnancy and odds of miscarriage, using a dataset in which cytokines were measured in pools of size 1 and 2, and in which replicates are indeed available. We include the discriminant function approach of Lyles *et al.* (2015) throughout, modified slightly to accommodate replicates, and provide accessible software for implementing both methods.

2.2 Methods

2.2.1 Poolwise logistic regression

We are interested in estimating parameters for an individual-level logistic regression model relating a binary outcome Y to a continuous exposure X and covariates $\boldsymbol{C}$:

$$\text{logit}[P(Y_{ij} = 1)] = \beta_0 + \beta_x X_{ij} + \boldsymbol{\beta_c^T C}_{ij} \tag{2.26}$$

Here i indexes the eventual pool number ($i = 1, ..., k$) and j indexes membership within a pool ($j = 1, .., g_i$), so Y_{ij} is the case status for the j^{th} member of the i^{th} pool comprised of g_i members ($g_i \in 1, 2, ...$). We consider a design in which each pool is homogeneous with respect to case status, i.e. comprised of either all cases ($Y_i = 1$) or all controls ($Y_i = 0$). This requires observing individual outcomes prior to forming pools in which to measure the biomarker.

Rather than observe individual-level biomarker levels for each member of the i^{th} pool, $\boldsymbol{X}_i = (X_{i1}, ..., X_{ig_i})^T$, we obtain from the assay a measure of the poolwise mean $\bar{X}_i = \frac{1}{g_i}\sum_{j=1}^{g_i} X_{ij}$, from which the poolwise sum can be calculated as $X_i^* = g_i\bar{X}_i$ (asterisks used to represent poolwise sums throughout). Individual-level covariate values $\boldsymbol{C}_i = (\boldsymbol{C}_{i1}, ..., \boldsymbol{C}_{ig_i})^T$ would typically be available, but we similarly calculate poolwise sums $\boldsymbol{C}_i^* = \sum_{j=1}^{g_i} \boldsymbol{C}_{ij}$.

In a case-control setting with no processing error or measurement error, Weinberg and Umbach (1999, 2014) provide the appropriate logistic regression model for estimating $\boldsymbol{\beta} = (\beta_0, \beta_x, \boldsymbol{\beta}_{\boldsymbol{c}}^T)^T$ based on the pooled variables:

$$\text{logit}[P(Y_i = 1)] = q_i + g_i\beta_0 + \beta_x X_i^* + \boldsymbol{\beta}_{\boldsymbol{c}}^T \boldsymbol{C}_i^* \tag{2.27}$$

with the offset q_i defined as:

$$q_i = g_i \text{log}\left(\frac{P(A|D)}{P(A|\bar{D})}\right) + g_i \text{log}\left(\frac{n_{\bar{D}}}{n_D}\right) + \text{log}\left(\frac{\#\text{ case pools of size } g_i}{\#\text{ control pools of size } g_i}\right) \tag{2.28}$$

where $P(A|D)$ and $P(A|\bar{D})$ are accrual probabilities for cases and controls and n_D and $n_{\bar{D}}$ are the total number of cases and controls across all pools. We note that if the disease prevalence p is known, the formula simplifies slightly. The accrual probability for cases is the number of cases sampled (n_D) divided by the number of cases in the population (N_D), and similarly for controls, so the first term in Eq. 2.28 becomes $g_i \text{log}\left(\frac{n_D/N_D}{n_{\bar{D}}/N_{\bar{D}}}\right)$, leading to

the offset formula:

$$q_i = g_i \log\left(\frac{1-p}{p}\right) + \log\left(\frac{\#\text{ case pools of size } g_i}{\#\text{ control pools of size } g_i}\right) \tag{2.29}$$

If accrual probabilities and disease prevalence are unknown, one can use Eq. 2.28 with the first term omitted. This only affects $\hat{\beta}_0$, which will typically be biased if there is case oversampling.

2.2.2 ML for handling errors in X_i^*

Following Schisterman *et al.* (2010), we assume the measurement obtained from the assay is not the precise poolwise mean $\bar{X}_i$, but the precise poolwise mean plus a processing error ϵ_i^p (if $g_i > 1$) plus a measurement error ϵ_i^m. Letting $\tilde{\bar{X}}_i$ represent the error-prone measurement, we can write:

$$\tilde{\bar{X}}_i = \bar{X}_i + \epsilon_i^p I(g_i > 1) + \epsilon_i^m \tag{2.30}$$

The poolwise logistic regression model Eq. 2.27 uses poolwise sums rather than means, which can be calculated as $\tilde{X}_i^* = g_i \tilde{\bar{X}}_i$.

In the i^{th} pool, we observe $(Y_i, \tilde{X}_i^*, \boldsymbol{C}_i^*)$. We write the likelihood contribution as:

$$L_i(\boldsymbol{\theta}) \propto f(Y_i, \tilde{X}_i^* | \boldsymbol{C}_i^*) = \int_{X_i^*} f(Y_i, \tilde{X}_i^*, X_i^* | \boldsymbol{C}_i^*)\, dX_i^* \tag{2.31}$$

Taking a classical measurement error modeling approach (Carroll *et al.*, 2006), a convenient factorization is:

$$\begin{aligned} L_i(\boldsymbol{\theta}) &= \int_{X_i^*} f(Y_i | \tilde{X}_i^*, X_i^*, \boldsymbol{C}_i^*)\; f(\tilde{X}_i^* | X_i^*, \boldsymbol{C}_i^*)\; f(X_i^* | \boldsymbol{C}_i^*)\, dX_i^* \\ &= \int_{X_i^*} f(Y_i | X_i^*, \boldsymbol{C}_i^*)\; f(\tilde{X}_i^* | X_i^*)\; f(X_i^* | \boldsymbol{C}_i^*)\, dX_i^* \end{aligned} \tag{2.32}$$

The simplification $f(Y_i|\tilde{X}_i^*, X_i^*, \boldsymbol{C}_i^*) = f(Y_i|X_i^*, \boldsymbol{C}_i^*)$ reflects a standard non-differential error assumption: the imprecise $\tilde{X}_i^*$ does not inform the outcome given the precise X_i^* and covariates (Carroll *et al.*, 2006). The result $f(\tilde{X}_i^*|X_i^*, \boldsymbol{C}_i^*) = f(\tilde{X}_i^*|X_i^*)$ reflects an assumption that the errors in Eq. 2.30 are independent of covariate values.

The first term under the integral in Eq. 2.32 is already specified by Eq. 2.27. For $\tilde{X}_i^*|X_i^*$, if we assume $\epsilon_i^p \sim N(0, \sigma_p^2)$ and $\epsilon_i^m \sim N(0, \sigma_m^2)$ and these errors are independent, then by Eq. 2.30 we have $\tilde{X}_i^* = g_i\tilde{\tilde{X}}_i = X_i^* + g_i\epsilon_i^p I(g_i > 1) + g_i\epsilon_i^m$, leading to:

$$\tilde{X}_i^*|X_i^* \sim N\left(X_i^*,\ g_i^2\sigma_p^2 I(g_i > 1) + g_i^2\sigma_m^2\right) \tag{2.33}$$

For $X_i^*|\boldsymbol{C}_i^*$, we first specify an individual-level model for $X_{ij}|\boldsymbol{C}_{ij}$ and then derive the implied poolwise model. A normal linear regression is common in the measurement error literature (Carroll *et al.*, 2006) and convenient here because it leads to a simple poolwise model. If we assume $X_{ij} = \alpha_0 + \boldsymbol{\alpha_c}^T\boldsymbol{C}_{ij} + \epsilon_{ij}^x,\ \epsilon_{ij}^x \overset{iid}{\sim} N(0, \sigma_x^2)$, then $X_i^* = \sum_{j=1}^{g_i} X_{ij} = g_i\alpha_0 + \boldsymbol{\alpha_c}^T\boldsymbol{C}_i^* + \epsilon_i^x,\ \epsilon_i^x \overset{ind}{\sim} N(0, g_i\sigma_x^2)$. Assuming ϵ_i^x is independent of ϵ_i^p and ϵ_i^m, the third term in Eq. 2.32 is:

$$X_i^*|\boldsymbol{C}_i^* \sim N(g_i\alpha_0 + \boldsymbol{\alpha_c}^T\boldsymbol{C}_i^*,\ g_i\sigma_x^2) \tag{2.34}$$

With the likelihood fully specified, optimization routines can be used to obtain ML estimates for $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T, \sigma_x^2, \sigma_p^2, \sigma_m^2)^T$. A variance-covariance matrix can be obtained by numerically approximating the Hessian at $\hat{\boldsymbol{\theta}}$ and taking its inverse. Both steps require integrating out X_i^* for each pool at each iteration.

2.2.3 Approximate ML

While numerically integrating the scalar X_i^*'s out of the likelihood function is feasible, we also consider a closed-form approximation that will be faster and perhaps more stable. Factoring

the likelihood slightly differently leads to an alternative to Eq. 2.32 for $L_i(\boldsymbol{\theta})$:

$$L_i(\boldsymbol{\theta}) = \left[\int_{X_i^*} f(Y_i|X_i^*, \boldsymbol{C}_i^*)\ f(X_i^*|\tilde{X}_i^*, \boldsymbol{C}_i^*)\ dX_i^*\right] f(\tilde{X}_i^*|\boldsymbol{C}_i^*) \tag{2.35}$$

The first density under the integral is specified by Eq. 2.27. For the second and third, we first derive the joint density $f(X_i^*, \tilde{X}_i^*|\boldsymbol{C}_i^*)$. Conditioning on $\boldsymbol{C}_i^*$ and using the poolwise linear regression Eq. 2.34, we can write:

$$\begin{bmatrix} X_i^* \\ \tilde{X}_i^* \end{bmatrix} = \begin{bmatrix} g_i\alpha_0 + \boldsymbol{\alpha_c}^T\boldsymbol{C}_i^* \\ g_i\alpha_0 + \boldsymbol{\alpha_c}^T\boldsymbol{C}_i^* \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 1 & g_iI(g_i > 1) & g_i \end{bmatrix} \begin{bmatrix} \epsilon_i^x \\ \epsilon_i^p \\ \epsilon_i^m \end{bmatrix} \tag{2.36}$$

Given the prior normality and independence assumptions, the error vector $\boldsymbol{\epsilon}_i = (\epsilon_i^x, \epsilon_i^p, \epsilon_i^m)^T$ is trivariate normal. Multivariate normal theory (Seber and Lee, 2012) dictates that $(X_i^*, \tilde{X}_i^*)^T$ is bivariate normal:

$$\begin{bmatrix} X_i^* \\ \tilde{X}_i^* \end{bmatrix} \sim N_2\left(\begin{bmatrix} g_i\alpha_0 + \boldsymbol{\alpha_c}^T\boldsymbol{C}_i^* \\ g_i\alpha_0 + \boldsymbol{\alpha_c}^T\boldsymbol{C}_i^* \end{bmatrix}, \begin{bmatrix} g_i\sigma_x^2 & g_i\sigma_x^2 \\ g_i\sigma_x^2 & g_i\sigma_x^2 + g_i^2I(g_i > 1)\sigma_p^2 + g_i^2\sigma_m^2 \end{bmatrix}\right) \tag{2.37}$$

This leads to two useful results. For the term outside the integral in Eq. 2.35:

$$\tilde{X}_i^*|\boldsymbol{C}_i^* \sim N\left(g_i\alpha_0 + \boldsymbol{\alpha_c}^T\boldsymbol{C}_i^*,\ g_i\sigma_x^2 + g_i^2I(g_i > 1)\sigma_p^2 + g_i^2\sigma_m^2\right) \tag{2.38}$$

And for the second term inside the integral:

$$X_i^*|(\tilde{X}_i^*, \boldsymbol{C}_i^*) \sim N(\bar{\mu}_i, \bar{\sigma}_i^2) \tag{2.39}$$

where $\bar{\mu}_i = \mu_{i1} + \frac{\Sigma_{i12}}{\Sigma_{i22}}(\tilde{X}_i^* - \mu_{i2})$ and $\bar{\sigma}_i^2 = \Sigma_{i11} - \frac{\Sigma_{i12}^2}{\Sigma_{i22}}$, with $(\mu_{i1}, \mu_{i2}, \Sigma_{i12}, \Sigma_{i22})$ apparent from Eq. 2.37.

With $Y_i|(X_i^*, \mathbf{C}_i^*) \sim \text{Bernoulli}\left(p_i = (1 + e^{-q_i - g_i\beta_0 - \beta_x X_i^* - \boldsymbol{\beta}_{\boldsymbol{c}}^T \boldsymbol{C}_i^*})^{-1}\right)$ and $X_i^*|(\tilde{X}_i^*, \boldsymbol{C}_i^*) \sim N(\bar{\mu}_i, \bar{\sigma}_i^2)$, the integral in Eq. 2.35 is a variant on a logistic-normal integral that arises in logistic regression with covariate measurement error outside of pooling. A closed-form approximation can be used to avoid numerical integration (Fuller, 1987; Lyles and Kupper, 2013). The first density under the integral in Eq. 2.35 is $p_i^{y_i}(1-p_i)^{1-y_i}$ where $p_i = H(\eta_i) = \frac{e^{\eta_i}}{1+e^{\eta_i}}$ and $\eta_i = q_i + g_i\beta_0 + \beta_x X_i^* + \boldsymbol{\beta}_{\boldsymbol{c}}^T \boldsymbol{C}_i^*$. Replacing the logistic function $H(\eta_i)$ with the probit approximation $\Phi(\frac{\eta_i}{k})$, where $\Phi(\cdot)$ is the standard normal CDF and typically $k = 1.7$ (Camilli, 1995), leads to:

$$p_i^* = P(Y_i = 1|\tilde{X}_i^*, \boldsymbol{C}_i^*) \approx H\left(\frac{q_i + g_i\beta_0 + \beta_x\bar{\mu}_i + \boldsymbol{\beta}_{\boldsymbol{c}}^T \boldsymbol{C}_i^*}{\sqrt{1 + \frac{\beta_x^2\bar{\sigma}_i^2}{1.7^2}}}\right) \tag{2.40}$$

Thus the closed-form expression $p_i^{*y_i}(1-p_i^*)^{1-y_i}$ can be used to approximate the integral in Eq. 2.35, which represents $f(Y_i|\tilde{X}_i^*, \boldsymbol{C}_i^*)$. Point estimates and standard errors can be obtained using the same procedures as for full ML.

2.2.4 Discriminant function approach

Lyles *et al.* (2015) developed an alternative to poolwise logistic regression for estimating the log-OR of interest while accounting for errors. The basic idea is to estimate the covariate-adjusted log-OR by fitting a normal-errors linear regression of X on $(Y, \boldsymbol{C})$ rather than a logistic regression of Y on $(X, \boldsymbol{C})$. The assumed individual-level model is:

$$X_{ij} = \gamma_0 + \gamma_y Y_{ij} + \boldsymbol{\gamma}_{\boldsymbol{c}}^T \boldsymbol{C}_{ij} + \epsilon_{ij},\ \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2) \tag{2.41}$$

It can be shown that the quantity $\frac{\gamma_y}{\sigma^2}$ corresponds to the same covariate-adjusted log-OR targeted by β_x in Eq. 2.27 (Lyles *et al.*, 2009). While the ML estimate is $\widehat{\text{log-OR}}_{ml} = \frac{\hat{\gamma}_y}{\hat{\sigma}^2}$, Lyles *et al.* (2015) provide a bias-adjusted version resulting from a second-order Taylor series expansion:

$$\widehat{\text{log-OR}}_{adj} = \widehat{\text{log-OR}}_{ml} - \frac{\hat{\gamma}_y \hat{V}(\hat{\sigma}^2)}{(\hat{\sigma}^2)^3} \tag{2.42}$$

If Eq. 2.41 holds true for individual data, then the corresponding poolwise model for $X_i^* | \boldsymbol{C}_i^*$ is:

$$X_i^* = \sum_{j=1}^{g_i} X_{ij} = g_i\gamma_0 + \gamma_y Y_i^* + \boldsymbol{\gamma_c}^T \boldsymbol{C}_i^* + \epsilon_i,\ \epsilon_i \overset{ind}{\sim} N(0, g_i\sigma^2) \tag{2.43}$$

In this case, Y_i^* is the number of subjects in the i^{th} pool with $Y_{ij} = 1$. This is different from the logistic regression setup where $Y_i = 1$ for case pools and 0 for control pools.

We again assume the assay returns the error-contaminated poolwise mean $\tilde{\bar{X}}_i = \bar{X}_i + \epsilon_i^p I(g_i > 1) + \epsilon_i^m$, from which the poolwise sum can be calculated as $\tilde{X}_i^* = g_i \tilde{\bar{X}}_i = X_i^* + g_i\epsilon_i^p I(g_i > 1) + g_i\epsilon_i^m$. The likelihood contribution for the observed $(Y_i^*, \tilde{X}_i^*, \boldsymbol{C}_i^*)$ is:

$$L_i(\boldsymbol{\theta}) \propto f(\tilde{X}_i^* | Y_i^*, \boldsymbol{C}_i^*) \tag{2.44}$$

where:

$$\tilde{X}_i^* | (Y_i^*, \boldsymbol{C}_i^*) \sim N\left(g_i\gamma_0 + \gamma_y Y_i^* + \boldsymbol{\gamma_c}^T \boldsymbol{C}_i^*,\ g_i\sigma^2 + g_i^2 I(g_i > 1)\sigma_p^2 + g_i^2\sigma_m^2\right) \tag{2.45}$$

Our result differs in a small but important way from that of Lyles *et al.* (2015). They assume that errors add to the poolwise sum X_i^*, while we consider it more plausible that they add to the poolwise mean $\bar{X}_i$ targeted by the assay.

The likelihood can be maximized to obtain $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\gamma}}^T, \hat{\sigma}^2, \hat{\sigma}_p^2, \hat{\sigma}_m^2)^T$, and $\hat{V}(\hat{\boldsymbol{\theta}})$ calculated as the inverse of the estimated Hessian at $\hat{\boldsymbol{\theta}}$. The bias-adjusted log-OR can then be calculated using Eq. 2.42. A delta method-based variance estimate for the MLE is $\hat{V}(\widehat{\text{log-OR}}_{ml}) =$

$f'(\hat{\boldsymbol{\theta}})\hat{V}(\hat{\boldsymbol{\theta}})f'(\hat{\boldsymbol{\theta}})^T$ with $f'(\hat{\boldsymbol{\theta}}) = (\frac{1}{\hat{\sigma}^2}, -\frac{\hat{\gamma}_y}{\hat{\sigma}^4})$. We use this variance estimator as an approximation for $\hat{V}(\widehat{\text{log-OR}})$; it should tend to be slightly conservative when used in conjunction with the bias-corrected estimator.

2.2.5 Incorporating replicates

For both logistic regression and the discriminant function approach, all parameters are identifiable without replicates provided there are a sufficient number of unique pool sizes. There must be at least two different pool sizes to correct for measurement error or processing error, and at least three different pool sizes including 1 to correct for both. But theoretical identifiability does not necessarily imply stable estimation (Carroll *et al.*, 2006). A relatively small number of replicate measurements may help to distinguish the variance components, particularly σ_m^2. While replicates for pools of any size could be accommodated, we focus primarily on replicate singles, where multiple assay measurements are obtained for some single-specimen pools.

Note that obtaining multiple assays for the same subjects requires a greater volume of biospecimen. This may not be feasible in certain scenarios, especially if pooling is utilized to reach the minimum volume for a particular assay. Replicates would still not be completely out of the question in these cases; whatever the minimum pool size is given available specimen volumes, one could form pools twice that size and have sufficient volume for two assays per pool. This is somewhat beyond our scope, and it is unclear whether two assays for the same pool would be subject to the same or different processing errors. We return to replicate singles.

If for the i^{th} single ($g_i = 1$) we obtain k_i independent assay measurements, $\tilde{\boldsymbol{X}}_i = (\tilde{X}_{i1}, ..., \tilde{X}_{ik_i})^T$, the logistic regression likelihood contribution for the i^{th} pool is the same as in Eq. 2.32 (asterisks omitted since $g_i = 1$), except $\tilde{\boldsymbol{X}}_i$ is vector-valued, so the second

term under the integral becomes $f(\tilde{\boldsymbol{X}}_i|X_i)$. There is no processing error for singles, so each $\tilde{X}_{ij}$ is the true biomarker level X_i plus an independent normally distributed measurement error. This can be written as: $\tilde{\boldsymbol{X}}_i = \mathbf{1}_{k_i}X_i + \boldsymbol{\epsilon}_i^m,\ \boldsymbol{\epsilon}_i^m \overset{ind}{\sim} N_{k_i}(\mathbf{0}_{k_i},\ \sigma_m^2\boldsymbol{I}_{k_i})$. It follows that:

$$\tilde{\boldsymbol{X}}_i|X_i \sim N_{k_i}(\mathbf{1}_{k_i}X_i,\ \sigma_m^2\boldsymbol{I}_{k_i}) \tag{2.46}$$

To incorporate replicates for approximate ML, we replace $\tilde{X}_i^*$ in Eq. 2.36 with $\tilde{\boldsymbol{X}}_i = \mathbf{1}_{k_i}(\alpha_0 + \boldsymbol{\alpha}_{\boldsymbol{c}}^T\boldsymbol{C}_i + \epsilon_i^x) + \boldsymbol{\epsilon}_i^m$, leading to a slightly modified version of Eq. 2.37 for $(X_i, \tilde{\boldsymbol{X}}_i)$ and subsequent results for $X_i|(\tilde{\boldsymbol{X}}_i, \boldsymbol{C}_i)$ and $\tilde{\boldsymbol{X}}_i|\boldsymbol{C}_i$.

For the discriminant function approach, the likelihood for a single-specimen pool with k_i replicates is $L_i(\boldsymbol{\theta}) = f(\tilde{\boldsymbol{X}}_i|Y_i, \boldsymbol{C}_i)$. The vector of replicates $\tilde{\boldsymbol{X}}_i$ can be written:

$$\begin{aligned}
\tilde{\boldsymbol{X}}_i &= \mathbf{1}_{k_i}X_i + \boldsymbol{\epsilon}_i^m \\
&= \mathbf{1}_{k_i}(\gamma_0 + \gamma_y Y_i + \boldsymbol{\gamma}_{\boldsymbol{c}}^T\boldsymbol{C}_i + \epsilon_i) + \boldsymbol{\epsilon}_i^m \\
&= \mathbf{1}_{k_i}(\gamma_0 + \gamma_y Y_i + \boldsymbol{\gamma}_{\boldsymbol{c}}^T\boldsymbol{C}_i) + \begin{bmatrix} \mathbf{1}_{k_i} & \boldsymbol{I}_{k_i} \end{bmatrix} \begin{bmatrix} \epsilon_i \\ \boldsymbol{\epsilon}_i^m \end{bmatrix}
\end{aligned} \tag{2.47}$$

The error vector $\boldsymbol{\epsilon}_i = (\epsilon_i, \boldsymbol{\epsilon}_i^{mT})^T$ is multivariate normal with mean $\mathbf{0}_{k_i+1}$ and a diagonal variance-covariance matrix with σ^2 and k_i σ_m^2's on the diagonal. Therefore:

$$\tilde{\boldsymbol{X}}_i|(Y_i, \boldsymbol{C}_i) \sim N_{k_i}\left(\mathbf{1}_{k_i}(\gamma_0 + \gamma_y Y_i + \boldsymbol{\gamma}_{\boldsymbol{c}}^T\boldsymbol{C}_i),\ \sigma^2\boldsymbol{J}_{k_i} + \sigma_m^2\boldsymbol{I}_{k_i}\right) \tag{2.48}$$

2.2.6 Implementation

We used R 3.5.0 (R Core Team, 2015) to develop the package **pooling** (Van Domelen, 2018b), which is available on GitHub and CRAN.

The functions *p_logreg_xerrors* and *p_dfa_xerrors* implement the logistic regression and

discriminant function approaches described here. User inputs include the poolwise data, whether to correct for measurement error, processing error, neither, or both, and other settings such as starting values and lower/upper bounds for parameters. Outputs include the estimated parameters, a variance-covariance matrix, and Akaike information criterion (Akaike, 1974).

We use the *nlminb* function in base R to maximize log-likelihood functions. Initial values are set to 0.01 for regression coefficients and 1 for variance components; lower bounds of 0.001 are used for the latter. Hessian matrices at the MLE's are numerically approximated using the *hessian* function from the **pracma** package v. 2.1.1 (Borchers, 2017). The logistic regression function supports both full ML and approximate ML. For full ML, the *hcubature* function in **cubature** v. 1.3-11 (Narasimhan and Johnson, 2017) is used for numerical integration.

2.3 Collaborative Perinatal Project

The Collaborative Perinatal Project (CPP) was a multisite prospective study initiated in 1959 and aimed at identifying risk factors for maternal and infant mortality and cerebral palsy (Hardy, 2003). A nested case-control study was later conducted to test whether serum cytokine levels measured during pregnancy were associated with risk of spontaneous abortion (SA) (Whitcomb *et al.*, 2007). We use data from the follow-up study, in which cytokines were measured in pools of size 1 and 2 using stored samples from the original study. Our research question is whether the cytokine monocyte chemotactic protein (*MCP-1*) is associated with risk of SA after adjusting for age, race, and smoking.

Data are comprised of 96 singles without replicates ($g_i = 1, k_i = 1$), 30 singles with two replicates ($g_i = 1, k_i = 2$), and 280 pools of size 2 ($g_i = 2, k_i = 1$), for a total of 686 subjects and 436 measurements.

Before applying the corrective methods, which rely on various assumed models and distributions, we note that the mean *MCP-1* was similar for cases and controls within each pool size. For $g = 1$, the mean (SD) was 0.131 (0.106) for cases and 0.124 (0.137) for controls (unequal variance t-test: p = 0.72); for $g = 2$, the mean (SD) for assay measurements (not multiplied by 2) were 0.226 (0.284) for cases and 0.203 (0.233) for controls (p = 0.48). While not adjusted for covariates, these comparisons are useful in that they provide estimates of the crude association between *MCP-1* and SA that rely on minimal distributional assumptions and are valid despite the errors (Abrevaya and Hausman, 2004; Carroll *et al.*, 2006).

Table 2.9 shows covariate-adjusted log-OR estimates for the corrective methods, with and without incorporating the 30 replicates. For the no-replicates analysis, one of the two *MCP-1* measurements was randomly selected for subjects with two measurements. Abbreviations are as follows: LRF = logistic regression, full ML; LRA = logistic regression, approximate ML; and DFA = discriminant function approach. Cell values indicate $\widehat{\text{log-OR}}$ (SE), AIC. *MCP-1* values were multiplied by 10 prior to fitting the models, so the $\widehat{\text{log-OR}}$'s are for a 0.1-ng/mL increment. Lower AIC values indicate better fits relative to models in the same row (Akaike, 1974).

Without replicates, either processing error or measurement error could be accounted for, but not both. Identifiability would require a third pool size in addition to 1 and 2. This is an important limitation because it means choosing from three candidate models that may all inadequately correct for *MCP-1* errors. AIC favored processing error only for all three corrective methods. Measurement error only models had lower AIC than neither, but produced implausible parameter estimates (e.g. residual error variances hitting 0.001). Logistic regression was particularly unstable for measurement error only, as different starting values produced very different $\widehat{\text{log-OR}}$'s but similar maximized log-likelihoods. Relative to neither error, processing only models had larger point estimates but also larger standard errors, such that the association between *MCP-1* and SA still did not approach significance.

Table 2.9: Estimates of adjusted log-OR for *MCP-1* and spontaneous abortion in CPP. Values are $\widehat{\text{log-OR}}$ (SE), AIC.

	Error type			
	Neither	PE only	ME only	Both
Without replicates				
LRF	0.012 (0.024), 2822.0	0.070 (0.114), 2697.6	-0.071 (-), 2762.4[b]	Not identifiable
LRA	n/a[a]	0.071 (0.115), 2697.5	-0.088 (-), 2762.4	Not identifiable
DFA	0.016 (0.025), 2277.6	0.090 (0.114), 2153.0	Inf (-), 2217.7[c]	Not identifiable
With replicates				
LRF	n/a[d]	n/a[d]	0.026 (0.049), 2353.8	0.046 (0.082), 2340.8
LRA	n/a[d]	n/a[d]	0.026 (0.049), 2353.8	0.046 (0.082), 2340.8
DFA	n/a[d]	n/a[d]	0.030 (0.051), 1809.5	0.050 (0.081), 1796.5

[a] No integral to approximate.

[b] $\hat{\sigma}_x^2 = 0.001$. SE omitted because variance-covariance matrix not positive definite.

[c] $\hat{\sigma}^2 = 0.001$, causing blow-up in $\widehat{\text{log-OR}} = \hat{\gamma}_y/\hat{\sigma}^2$.

[d] Non-identical replicates are incompatible with no measurement error.

With replicates incorporated, the two candidate models are measurement error only and both errors, since non-identical replicates are incompatible with no measurement error (in Eq. 2.46, $\sigma_m^2 = 0$ implies $\tilde{X}_{i1} = \tilde{X}_{i2} = X_i$). AIC favored both errors for all three methods. The estimated variance components for LRF were as follows: $\hat{\sigma}_x^2 = 1.580$, $\hat{\sigma}_p^2 = 0.729$, and $\hat{\sigma}_m^2 = 0.108$. Relatively small measurement error variance is reasonable given the high correlation for the 30 *MCP-1* replicates ($r = 0.976$).

Table 2.10 compares the LRF fit accounting for both error types alongside the naive poolwise logistic regression fit treating *MCP-1* values as precise poolwise sums. Both models suggest that older age, non-white race, and current smoking are associated with higher odds of SA. The covariate-adjusted association for *MCP-1* and SA was not statistically significant in either model, but the estimated OR was slightly higher in the error-adjusted model.

Table 2.10: Logistic regression estimates for odds of spontaneous abortion in CPP.

	Ignoring MCP-1 errors		*Accounting for MCP-1 errors*	
Variable	Beta (SE)	OR (95% CI)	Beta (SE)	OR (95% CI)
Intercept	-1.565 (0.372)	-	-1.581 (0.374)	-
MCP-1	0.012 (0.024)	1.012 (0.966, 1.060)	0.046 (0.082)	1.047 (0.891, 1.230)
Mother's age	0.037 (0.013)	1.037 (1.011, 1.064)	0.036 (0.013)	1.037 (1.011, 1.064)
Non-white race	0.560 (0.175)	1.751 (1.242, 2.470)	0.566 (0.176)	1.761 (1.247, 2.488)
Current smoking	0.338 (0.162)	1.402 (1.021, 1.926)	0.338 (0.162)	1.402 (1.021, 1.926)

2.4 Simulation studies

We performed simulations modeled after the CPP data to assess validity and efficiency of our methods and to explore the effect of errors on the relative efficiency of pooling designs vs. traditional.

The main assumption underlying the discriminant function approach ($X|(Y, \boldsymbol{C})$ is a normal-errors linear regression) implies a logistic regression model for $Y|(X, \boldsymbol{C})$, while the assumptions underlying the logistic regression method (homogeneous pools, logistic regression for $Y|(X, \boldsymbol{C})$, linear regression for $X|\boldsymbol{C}$) do not necessarily imply the discriminant function model. Our main focus is logistic regression, where odds ratios for all predictors can be estimated, rather than just for the pooled biomarker. So we generate data under logistic regression and compare parameter estimation for both methods, where the discriminant function approach is more of a working model for the data.

Covariates generated independently of each other include mother's age, $C_{1ij} \in (14, ..., 45)$ with sampling probabilities matching the CPP age distribution; non-white race, $C_{2ij} \sim$ Bernoulli(0.34); and smoking, $C_{3ij} \sim$ Bernoulli(0.47). Using estimates from the full-ML logistic regression with both error types and replicates, *MCP-1* in 10 ng/mL (X_{ij}) given $\boldsymbol{C}_{ij}$ is a linear regression with $(\alpha_0, \boldsymbol{\alpha}_{\boldsymbol{c}}^T, \sigma_x^2) = (0.50, 0.03, -0.17, 0.02, 1.58)$, and SA ($Y_{ij}$)

given $(X_{ij}, \boldsymbol{C}_{ij})$ is a logistic regression with $(\beta_0, \beta_x, \boldsymbol{\beta}_{\boldsymbol{c}}^T) = (-1.58, 0.20, 0.04, 0.57, 0.34)$. The estimated log-OR for *MCP-1* was 0.046, but we use 0.20 to simulate a moderate effect where a 0.1-ng/mL increment in *MCP-1* is similar to a 5-year increment in mother's age in terms of its impact on odds of SA. Finally, error variances were set to $\sigma_p^2 = 0.729$ and $\sigma_m^2 = 0.108$.

2.4.1 Validity of error-correction methods

The first set of simulations is aimed at assessing validity under a hybrid design with pools of size 1, 2, and 3, where all parameters are identifiable even without replicates. For each trial, we generate 686 values for $(\boldsymbol{C}_{ij}, X_{ij}, Y_{ij})$ and separate the data into n_1 cases and n_0 controls. Within cases, we form $\frac{n_1}{6}$ (rounded up) pools of size 2 and 3 and leave the remaining observations as singles, and similarly for controls. For the pooled biomarker, we calculate the true poolwise mean $\bar{X}_i$, add normal errors to obtain the imprecise poolwise mean $\tilde{\bar{X}}_i$, and multiply by the pool size to obtain the imprecise poolwise sum $\tilde{X}_i^*$. For scenarios with replicates, $\tilde{X}_i = (\tilde{X}_{i1}, \tilde{X}_{i2})^T$ is generated by adding two independent measurement errors to X_i.

Table 2.11 summarizes performance of the three methods alongside naive poolwise logistic regression ignoring errors under processing error only, measurement error only, and both. Error type here refers to both data generation and estimation, such that LRF, LRA, and DFA are always based on correctly specified errors.

In the processing error only scenario, the naive logistic regression ignoring errors exhibited substantial downward bias and poor CI coverage, suggesting the processing errors are too large to ignore. The corrective methods performed reasonably well, although LRF and LRA had some upward bias. Despite generating data under logistic regression, DFA had slightly less bias and better efficiency than LRF and LRA.

For measurement error only, there was only a small amount of downward bias and slightly

Table 2.11: Simulation results for estimation of adjusted log-OR for *MCP-1* and spontaneous abortion (2,500 trials each, true log-OR = 0.20).

	Mean bias (median bias)	SD (IQR)	Mean SE	MSE	Coverage
Processing error only					
Naive	-0.097	0.047	0.047	0.012	0.447
LRF	0.014	0.099	0.099	0.010	0.958
LRA	0.013	0.099	0.098	0.011	0.958
DFA	0.005	0.092	0.094	0.009	0.959
Measurement error only					
Without replicates[a]					
Naive	-0.025	0.064	0.062	0.005	0.919
LRF	0.027	0.108	0.104	0.012	0.970
LRA	0.027	0.107	0.104	0.012	0.970
DFA	0.003	0.084	0.095	0.007	0.970
With replicates					
LRF	0.005	0.076	0.074	0.006	0.954
LRA	0.004	0.076	0.074	0.006	0.954
DFA	0.001	0.074	0.073	0.005	0.954
Both error types					
Without replicates[b]					
Naive	(-0.106)	(0.061)	-	-	0.355
LRF	(0.062)	(0.306)	-	-	0.987[d]
LRA	(0.062)	(0.302)	-	-	0.987[e]
DFA[c]	(0.053)	(0.290)	-	-	0.986
With replicates					
LRF	0.014	0.103	0.103	0.011	0.962
LRA	0.013	0.102	0.102	0.011	0.962
DFA	0.005	0.095	0.098	0.009	0.964

[a] Excludes 2 trials in which $\widehat{\text{log-OR}} > 2$ for LRF.

[b] Median bias and IQR reported to lessen impact of extreme estimates.

[c] Bias adjustment not used because it frequently flipped sign of $\widehat{\text{log-OR}}$.

[d] Excludes 70 trials in which variance-covariance matrix was not positive definite.

[e] Excludes 79 trials in which variance-covariance matrix was not positive definite.

lower than nominal coverage for the naive method, suggesting the measurement errors were nearly small enough to ignore. Without replicates, LRF and LRA exhibited upward bias of about the same magnitude as the naive approach, while DFA was virtually unbiased and more efficient. The measurement error variance estimate $\hat{\sigma}^2_m$ hit its lower bound in 23.5% of trials for both LRF and LRA and 27.7% of trials for DFA. Replicates improved estimation; for all three methods, $\hat{\sigma}^2_m$ never hit 0.001, bias was reduced and efficiency improved, and CI coverage was closer to nominal.

In the both-errors scenario, performance without replicates was poor. The corrective methods often produced extreme estimates ($\widehat{\text{log-OR}}$ outside of [-1, 1] in 13.1% of trials for LRF, 12.6% for LRA, 13.9% for DFA) and exhibited upward median bias. At least one variance component estimate hit 0.001 in the majority of trials for all three methods (LRF: $\hat{\sigma}^2_x$ 0.1%, $\hat{\sigma}^2_p$ 16.8%, $\hat{\sigma}^2_m$ 48.0%; LRA: $\hat{\sigma}^2_x$ 0.1%, $\hat{\sigma}^2_p$ 16.9%, $\hat{\sigma}^2_m$ 47.8%; DFA: $\hat{\sigma}^2$ 4.8%, $\hat{\sigma}^2_p$ 20.3%, $\hat{\sigma}^2_m$ 48.2%). Adding replicates resolved this issue and drastically improved performance.

The stabilizing role of replicates in the both-errors scenario is illustrated by the $\widehat{\text{log-OR}}$ histograms in Figure 2.5. While the log-OR is identifiable with pools of size 1, 2, and 3 and no replicates, estimation is relatively unstable even for a fairly large sample size. We note that log-OR estimates outside of [-1, 1] remained fairly common even after a 5-fold increase in sample size to $n = 3,430$ (1,000 trials: 2.9% for LRF, 3.9% for LRA, 2.9% for DFA).

DFA was more efficient than LRF and LRA across all scenarios, and additional gains may be possible if $H_0 : \gamma_y = 0$ rather than $H_0 : \text{log-OR} = \frac{\gamma_y}{\sigma^2} = 0$ is used for the primary test of association. The former is indeed a uniformly most powerful unbiased test for $H_0 : \text{OR} = 1$ (Lyles *et al.*, 2009). To briefly explore this, we ran an additional 2,500 trials of the processing error only scenario from Table 2.11. Empirical power was 0.588 for LRA, 0.612 for DFA with $H_0 : \text{log-OR} = 0$ (bias-adjusted version), and 0.637 for DFA with $H_0 : \gamma_y = 0$.

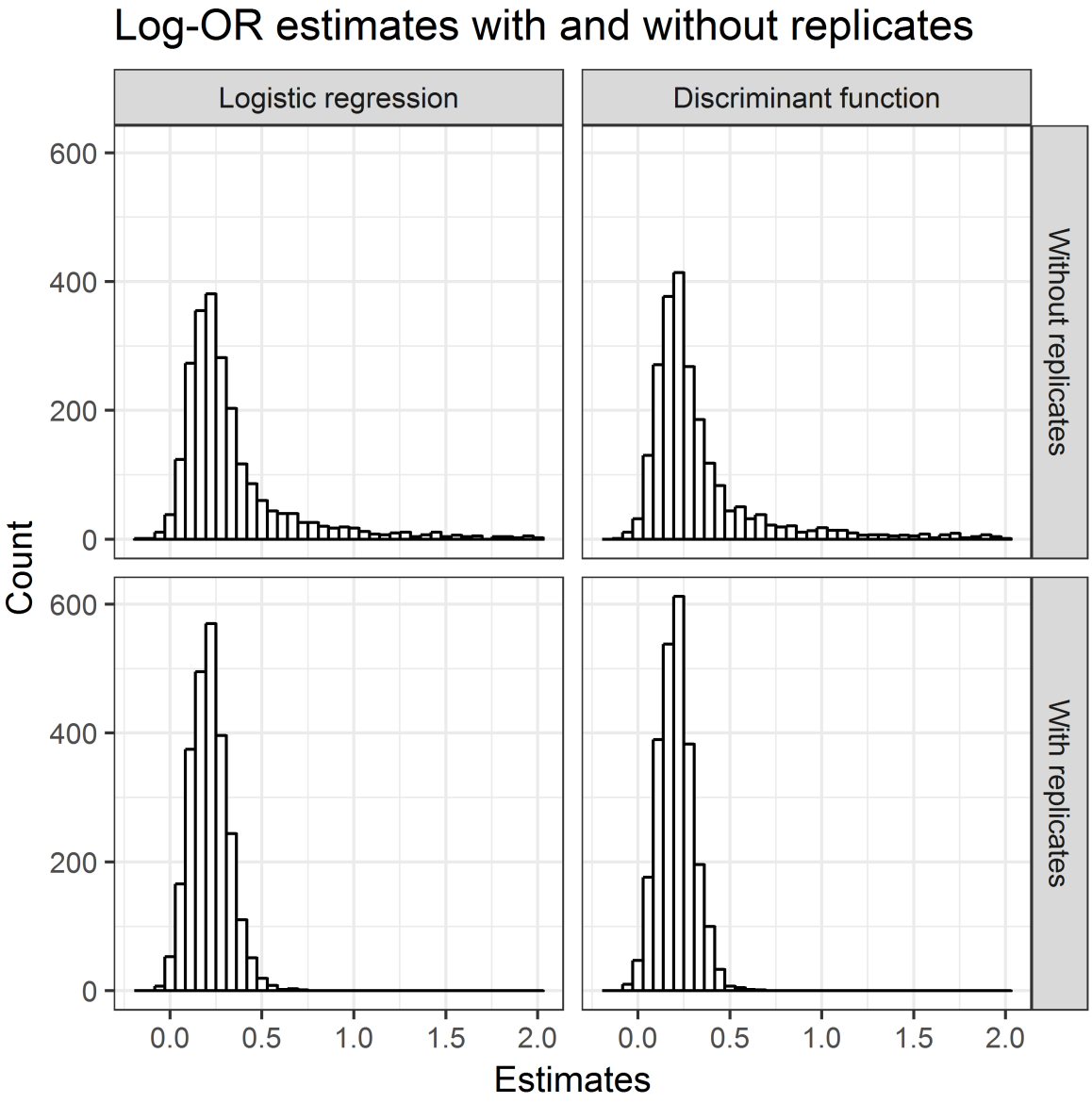

Figure 2.5: Histograms of log-OR estimates in simulations with processing error and measurement error (2,500 trials, true log-OR = 0.2).

2.4.2 Robustness to non-normality of errors

We re-ran the prior simulations with errors distributed lognormal rather than normal to assess robustness to non-normal errors. Processing errors were generated from LN$(0.925, 0.099)$ minus 2.6498 and measurement errors from LN$(-0.022, 0.099)$ minus 1.0279, corresponding to mean-0, skewness of 1, and variances of 0.73 and 0.11, respectively, as in the previous normal-errors scenario. Results are summarized in Table 2.12. All three methods performed well despite modeling right-skewed lognormal errors as normal; performance metrics were extremely similar to the normal-errors results in Table 2.11. Performance was also similar with errors uniformly distributed with mean 0 and variances 0.73 and 0.11 (not shown).

2.4.3 Efficiency of traditional vs. pooling designs

Next we compare the efficiency of various study designs holding the total number of assays fixed and varying the magnitude of processing errors and measurement errors. For each trial, we generate 50,000 $(\boldsymbol{C}_{ij}, X_{ij}, Y_{ij})$ values using the same parameters as in the previous simulations. For the traditional design, we sample 450 cases and 450 controls.

For the first pooling design ("P-1-2-3"), we sample 900 cases and form 450 pools—150 with $g = 1$, 150 with $g = 2$, and 150 with $g = 3$—and similarly for controls. For the second, more aggressive pooling design ("P-1-5"), we sample 1,650 cases and form 450 pools—150 with $g = 1$ and 300 with $g = 5$—and also for controls. In scenarios where there is measurement error, the traditional design requires replicates for identifiability, so we randomly select 50 observations for which to generate two exposure measurements and 50 to exclude to keep the assay count at 900. We also incorporate 50 replicate singles into both pooling designs in scenarios with measurement error, while dropping 50 non-replicate singles to maintain 900 total assays.

Table 2.12: Simulation results for estimation of adjusted log-OR for *MCP-1* and spontaneous abortion, with errors distributed lognormal rather than normal (2,500 trials each, true log-OR = 0.20).

	Mean bias (median bias)	SD (IQR)	Mean SE	MSE	Coverage
Processing error only					
Naive	-0.097	0.050	0.048	0.012	0.458
LRF	0.012	0.102	0.099	0.011	0.956
LRA	0.012	0.101	0.099	0.010	0.956
DFA	0.004	0.094	0.094	0.009	0.960
Measurement error only					
Without replicates					
Naive	-0.024	0.064	0.063	0.005	0.921
LRF	0.028	0.109	0.104	0.013	0.969
LRA	0.027	0.107	0.103	0.012	0.969
DFA	0.004	0.085	0.094	0.007	0.964
With replicates					
LRF	0.006	0.077	0.074	0.006	0.950
LRA	0.006	0.077	0.074	0.006	0.949
DFA	0.003	0.074	0.073	0.006	0.951
Both error types					
Without replicates					
Naive	(-0.106)	(0.062)	-	-	0.362
LRF	(0.067)	(0.366)	-	-	0.984[b]
LRA	(0.066)	(0.359)	-	-	0.985[c]
DFA[a]	(0.060)	(0.348)	-	-	0.984
With replicates					
LRF	0.016	0.109	0.104	0.012	0.959
LRA	0.015	0.108	0.103	0.012	0.959
DFA	0.007	0.100	0.099	0.010	0.956

[a] Bias adjustment not used because it frequently flipped sign of $\widehat{\log\text{-OR}}$.

[b] Excludes 88 trials in which variance-covariance matrix was not positive definite.

[c] Excludes 99 trials in which variance-covariance matrix was not positive definite.

Figure 2.6 compares efficiencies for the traditional and pooled designs for the LRA and DFA methods (LRF omitted; $r > 0.998$ for LRF and LRA in first 25 trials for all scenarios). Trends for pooling vs. traditional designs were similar for LRA (left column) and DFA (right). For processing error only, the pooling designs were highly efficient for small processing error, but that advantage eroded and eventually reversed as σ_p^2 increased. For measurement error only and both error types, the efficiency advantage was reduced with increasing measurement error, but the pooling designs did not become clearly counterproductive even for large σ_m^2. Notably, DFA was more efficient than LRA in 53 out of 54 scenarios.

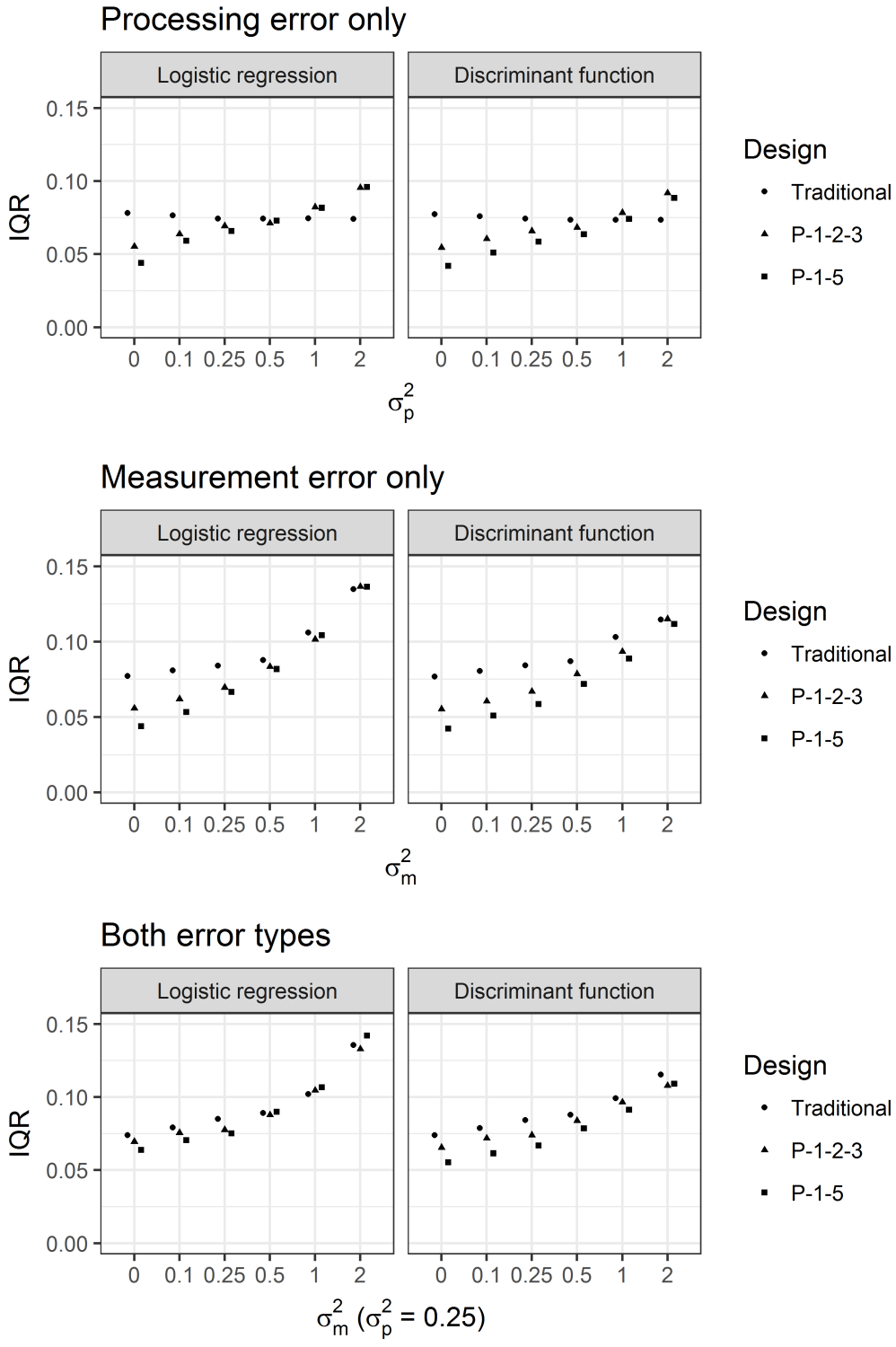

Figure 2.6: Interquartile range of log-OR estimates (5,000 trials each).

2.5 Discussion

Weinberg and Umbach (1999) developed a homogeneous pools logistic regression model that provides an analytic method to accompany a cost-effective pooling design, which can be used in any scenario where outcomes are observed prior to measuring exposure (e.g. cross-sectional and case-control studies, and cohort studies with stored specimens). However, fitting this model without accounting for potential errors in the poolwise biomarker measurements can lead to bias. Validity requires not only that the assay has negligible measurement error, but also that each value it returns is exactly the arithmetic mean exposure for members of a pool. In reality, handling and combining samples in the lab may lead to extra variability that cannot be ignored.

In general, the corrective methods we examined to correct for errors produced valid log-odds ratio estimates. Our updates to a proposed discriminant function approach (Lyles *et al.*, 2015) tended to give less biased and in some cases considerably more efficient estimates of the exposure log-odds ratio than the newly developed logistic regression approach in simulations, despite generating data under logistic regression. The bias adjustment incorporated into the discriminant function approach (see Eq. 2.42) likely explains some of this difference, as logistic regression is prone to small-sample bias away from the null (Nemes *et al.*, 2009; Firth, 1993). Nevertheless, we suspect analysts may still prefer logistic regression, given that it is the more familiar and general of the two and yields log-odds ratio estimates for all predictors rather than just the pooled biomarker. Both methods can theoretically correct for both processing error and measurement error as long as there are at least three different pool sizes including 1, but we find that adding replicate single measurements drastically improves stability when both errors are present.

For logistic regression, full and approximate ML produced very similar parameter estimates for the CPP dataset and had extremely similar performance in simulations. Full ML

is much slower because it requires numerical integration for each pool at each iteration of likelihood maximization. For our Table 2.11 simulations with both error types and replicates, each trial took approximately 5 minutes for full ML, and only about 3 seconds for approximate ML. In practice, investigators with poolwise data could fit both versions, confirm that estimates are similar, and report the full ML results. Comparing parameter estimates and maximized log-likelihoods might also be helpful in detecting numerical issues with full ML when they occur.

Our approaches are fully parametric and thus potentially susceptible to validity issues when assumptions are violated. Simulations suggested some robustness to non-normal errors, but the error distributions we tested were still mean-0 and additive. If normality assumptions are clearly violated, one could consider using our full ML logistic regression framework with a different measurement error model and/or exposure model. But alternative exposure models (e.g. a log-transformed linear regression) will typically not have a convenient poolwise sum result like linear regression (Mitchell *et al.*, 2014a). The discriminant function approach could also be used with non-normal errors, but it would likely not have a closed-form likelihood.

The discriminant function method has some appealing features in terms of study design considerations. First, it does not require homogeneous pools—perhaps a minor point considering that pooling-related efficiency gains require homogeneous pooling. Second, it holds a notable efficiency advantage over logistic regression, which might be even more pronounced if one were to base inference on the regression coefficient γ_y rather than the log-odds ratio $\frac{\gamma_y}{\sigma^2}$. As Lyles *et al.* (2009) point out, when the assumptions underlying the discriminant function model are met, $H_0 : \gamma_y = 0$ is a uniformly most powerful unbiased test for whether the odds ratio equals 1. We indeed observed a modest gain in empirical power from targeting γ_y rather than $\frac{\gamma_y}{\sigma^2}$ (0.637 vs. 0.612) in a processing error-only simulation. The empirical power for logistic regression was 0.588, so performing inference on γ_y boosted the power advantage of the discriminant function approach from 0.024 to 0.049.

On a related note, the idea of targeting γ_y via the discriminant function model motivates a compelling study design. One could use a single pool size (ideally a large one) and not worry about identifying each variance component. Processing and measurement errors would simply add to the residual error variance. Inference for γ_y would be valid, just with reduced power relative to processing and measurement errors not being present (Carroll *et al.*, 2006). The log-odds ratio itself would not be estimable, but one could use $\frac{\hat{\gamma}_y}{MSE}$ for a lower bound (the MSE would overestimate σ^2 due to the additional errors). A sensitivity analysis could be used to gauge plausible values for the true log-odds ratio. This design would be simple and likely very powerful, since there would be no need to use up assays on replicates or smaller pool sizes included solely to help distinguish variance terms.

Our methods assume that errors in the pooled biomarker have the same form for cases and controls. While assuming non-differential measurement error is dubious when exposure levels are self-reported (Carroll *et al.*, 2006; White, 2003), it seems reasonable for the assay-based assessment we are considering. It is unlikely that assay errors would differ by case status (differential measurement error) or that case samples and control samples might be handled in a way that induces different amounts of extra variability to each (differential processing error). The latter could perhaps occur in scenarios where new controls are matched to case samples that have been stored for an extended period of time.

One way to relax the non-differential error assumption is to allow the processing and/or measurement errors to have different variances in case pools and control pools. While this would nullify the discriminant function approach's advantage of not requiring homogeneous pools, the pooling design lacks an efficiency advantage in that design anyway (Lyles *et al.*, 2015). We have included options to allow for differential errors in our publicly available R functions.

A similar concern is whether it is reasonable to assume that the processing error variance is independent of pool size. If caused by factors such as unequal specimen volumes and

cross-reactions among samples from different subjects, it may be more pronounced in larger pools. We suggest two potential solutions. First, one can avoid the problem entirely with a design that includes singles and pools of just one other size, such as the P-1-5 design. The models discussed herein would account for whatever processing error affects the pooled observations; it would not matter whether pools of other sizes would have had larger or smaller errors. Second, one could specify a relationship between pool size and processing error variance. One simple approach currently supported in our R functions is to assume the assay returns the poolwise mean plus a normal processing error times $\sqrt{\frac{g_i}{2}}I(g_i > 1)$ (plus the measurement error, if applicable). This reflects an assumption that the processing error variance increases at the same rate as the pool size, so that for example a pool with 2x the number of members would be subject to processing error with 2x the variance. Other more flexible approaches are also possible, such as a linear relationship between pool size and processing error variance with a non-unity slope estimated from the data.

A brief note on identifiability in the absence of replicates is warranted, as our assessment differs slightly from those of prior authors (Lyles *et al.*, 2015; Schisterman *et al.*, 2010). Returning to the original set of assumptions (non-differential errors, processing error variance independent of pool size), the variance of the error-prone poolwise sum biomarker level given covariates is $g_i\sigma_x^2 + g_i^2\sigma_p^2 I(g_i > 1) + g_i^2\sigma_m^2$. With measurement error only, two pool sizes g_1 and g_2 result in variances $g_1\sigma_x^2 + g_1^2\sigma_m^2$ and $g_2\sigma_x^2 + g_2^2\sigma_m^2$, respectively. For any two distinct pool sizes (g_1, g_2), these quantities are not equal nor multiples of each other, so σ_x^2 and σ_m^2 are identified. The situation is the same with processing error only: at least two different pool sizes are required, and neither has to be 1 (Lyles *et al.*, 2015). With both error types, we agree that at least three different pool sizes including 1 are required to identify all parameters (Lyles *et al.*, 2015; Schisterman *et al.*, 2010). However, two pool sizes *not* including 1 are sufficient to identify σ_x^2 and the sum $(\sigma_p^2 + \sigma_m^2)$, which in theory is enough to achieve the primary goal of removing bias due to both error types. If replicate singles are included in

the study design, identifiability is guaranteed regardless of what pool sizes are included.

The fact that two pool sizes other than 1 is sufficient to correct for both error types, while not bothering to distinguish them, is initially encouraging. It suggests a way to get around stability issues that arise when both errors are present and there are no replicates. In this scenario, each poolwise measurement is subject to a normal processing error and a normal measurement error, which can be viewed as a single mean-0 normal error with variance $\sigma_p^2 + \sigma_m^2$. This is no different than a processing error only scenario with error variance $\sigma_p^2 + \sigma_m^2$, so we might expect similar stability. Unfortunately, adequate stability in processing error only scenarios is aided by the very presence of singles, which help to distinguish σ_p^2. In our second set of simulations, we experimented with a P-2-3 design, but found that it was even less stable than P-1-2-3 (e.g. 52% larger IQR in 1,000 trials with $\sigma_p^2 = \sigma_m^2 = 0.1$).

Next, we turn to the central question of whether pooling remains cost-effective in the presence of errors. In a two-sample t-test scenario, a pooling design where each measurement is the arithmetic mean for g members of a group is efficient because each measurement has variance $\frac{\sigma^2}{g}$ rather than σ^2. The ratio of variances for pooled measurements to individual measurements is $\frac{1}{g}$, so the optimal design for a fixed number of assays is one very large pool size. Theoretically, a large enough pool size could provide power of virtually 1 for any fixed number of assays.

With errors, the variance of each measurement in the traditional design is $\sigma^2 + \sigma_m^2$, and in the pooling design is $\frac{\sigma^2}{g_i} + \sigma_p^2 + \sigma_m^2$. The ratio is $V_{p:t} = \frac{1}{\sigma^2+\sigma_m^2}(\frac{\sigma^2}{g_i} + \sigma_p^2 + \sigma_m^2)$, which is minimized for $\sigma_p^2 = \sigma_m^2 = 0$. So processing error and measurement error both have the effect of reducing the efficiency advantage of a pooling design.

If there is processing error only, $V_{p:t} = \frac{1}{g_i} + \frac{\sigma_p^2}{\sigma^2}$, which converges to $\frac{1}{g_i}$ as $\sigma_p^2 \to 0$, ∞ as $\sigma_p^2 \to \infty$, and $\frac{1}{g_i} + 1$ as $\sigma_p^2 \to \sigma^2$. We note that $V_{p:t} > 1$ if $\sigma_p^2 > \sigma^2(1 - \frac{1}{g_i})$, meaning that, for example, if the biggest pool size possible is 5, a pooling design will be less efficient than a

traditional one if σ_p^2 is more than 80% of σ^2.

For measurement error only, $V_{p:t} = (\frac{\sigma^2}{g_i} + \sigma_m^2)/(\sigma^2 + \sigma_m^2)$ which converges to $\frac{1}{g_i}$ as $\sigma_m^2 \to 0$, 1 as $\sigma_m^2 \to \infty$, and $0.5 < \frac{1+g_i}{2g_i} < 1$ as $\sigma_m^2 \to \sigma^2$.

To summarize, for processing error only, a pooling design can become counterproductive if σ_p^2 is nearly as large or larger than σ^2, but for measurement error only, a pooling design should remain more efficient even if σ_m^2 is as large as σ^2. With both errors, results are generally the same as for processing error only, but the added measurement error will make any efficiency advantage smaller than it would have been with only processing error and the same σ_p^2.

While our analytic framework is somewhat different (i.e. the models are more involved, covariates are present, and the variance terms have to be estimated), our simulations (see Figure 2.6) mostly agreed with efficiency results predicted by the above t-test-based arguments.

In summary, we have provided a method to correct for errors that can compromise validity of homogeneous pools logistic regression. The pooling design should remain cost-effective in situations where the assay is expensive and relatively precise, and careful handling can keep processing errors to a minimum. In future work, we plan to further generalize the methods presented here to accommodate non-normal errors and skewness in the pooled biomarker. Developing methods to handle potential sources of bias in pooling studies should lead to more feasible implementation of this very promising study design.

Chapter 3: Gamma models to accommodate a skewed exposure measured in pools and subject to multiplicative errors

3.1 Introduction

In the logistic regression setting where measuring a continuous exposure requires an expensive assay, a pooling study design can be extremely cost-effective (Weinberg and Umbach, 1999; Mitchell *et al.*, 2014b,a; Lyles *et al.*, 2016). We consider a design in which the assay is applied to pooled rather than individual biospecimen samples, with each pooled sample comprised of an equal volume from some number of like participants with respect to case status (i.e. all cases or all controls). Assuming the assay returns the mean biomarker level for members of each pool, the logistic regression model provided by Weinberg and Umbach (1999, 2014) can be used to estimate the log-odds ratios of interest with poolwise sums rather than individual-level data.

However, two types of error may affect pooled biomarker measurements and induce bias if ignored. Measurement error is extra variability due to assay imprecision, and processing error

is extra variability due to physically combining biospecimens into pools (Schisterman *et al.*, 2010). In a hybrid design that includes some single-specimen pools, measurement error would impact all measurements, while processing error would only affect multi-specimen pools. While measurement error could be assumed non-existent or negligible in scenarios where the assay is known to be very accurate, it seems generally dubious to assume no processing error. That would require precise formation of exactly equal-volume pools and minimal changes in the pooled biomarker concentration caused by cross-reactions from mixing biospecimens from different subjects.

Following the framework of Schisterman *et al.* (2010), Lyles *et al.* (2015) used maximum likelihood to estimate the covariate-adjusted log-odds ratio for a pooled biomarker subject to processing error and measurement error. They used a discriminant function approach in which the exposure log-odds ratio is estimated not from a logistic regression, but from a linear regression of the exposure on case status and covariates. The primary assumptions were as follows: (1) exposure level given case status and covariates is normally distributed with homoscedastic errors; (2) measurement errors and processing errors are additive, independent, normally distributed with mean 0 and variances independent of pool size; and (3) measurement errors affect all measurements, while processing errors only affect pools of size 2 or larger. Advantages of this approach include its computational simplicity, its applicability to designs with homogeneous or heterogeneous pools with respect to case status, the availability of a small-sample bias correction, and the ability to correct for both error types without replicate assay measurements, provided there are at least three different pool sizes including pools of size 1. A notable disadvantage is that it produces an odds ratio estimate for the pooled biomarker, but not for covariates.

In Chapter 2, we relied on similar error assumptions as Schisterman *et al.* (2010) and Lyles *et al.* (2015) to correct for assay errors in fitting the Weinberg and Umbach (1999) poolwise logistic regression model. Taking a classical measurement error modeling approach

(Carroll *et al.*, 2006), we wrote the likelihood contribution for the i^{th} pool as the product of three densities—case status given biomarkers and covariates, error-prone biomarker given true biomarker, and true biomarker given covariates—with the unobserved true biomarker level integrated out.

An important limitation of the Lyles *et al.* (2015) approach and our logistic regression methods from Chapter 2 is that they fail to address two common features of biomarker distributions: skewness and positivity (Frerichs *et al.*, 1976; Mendall *et al.*, 1996). In this chapter, we address this limitation by providing Gamma-based analogues to both of these approaches.

In considering how to adapt our logistic regression approach, a natural idea is to assume a linear model for the log-transformed biomarker level given covariates, and perhaps lognormal multiplicative errors rather than normal additive errors. A linear model for log-biomarker level given covariates implies a linear model for the sum of the log-biomarker levels for members of a given pool vs. summed covariates. But the summed log-biomarker level for each pool cannot be recovered from the poolwise mean, which is what the assay is assumed to target.

We consider an alternative approach with a more convenient poolwise-sum result: a Gamma regression model for exposure given covariates, along with multiplicative lognormal errors. Following the "alternate Gamma model" approach from Mitchell *et al.* (2015), we assume that the biomarker given covariates is Gamma distributed with constant scale parameter and shape parameter log-linear in covariates. Under this setup, each covariate is linearly related to the log of the expected value of the biomarker, and the variance of the biomarker level is directly proportional to the expected value.

Assuming independence among members of a pool, the corresponding model for poolwise sum biomarker level given summed covariates is Gamma with the same scale parameter as for individual-observations and shape parameter the sum of the individual-level shape parame-

ters (Lehmann and Casella, 2006). Unlike the log-transformed linear regression, this result is compatible with observed data, provided individual-level covariates are available. We further assume multiplicative lognormal errors, using a generally similar set of assumptions as in previous work (Schisterman *et al.*, 2010; Lyles *et al.*, 2015). We accommodate replicate assay measurements, which are again not strictly required for identifiability but may help stabilize ML estimation. A notable feature of the Gamma setup is that the log-odds ratios are theoretically identifiable even if there are no replicates and only one pool size. This stems from the fact that observing the product of a Gamma variable and a lognormal variable permits estimating their separate parameters, while observing the sum of two normal variables does not.

To adapt the Lyles *et al.* (2015) discriminant function approach, we modify the framework of Whitcomb *et al.* (2012) to include covariates and errors. We use a similar Gamma model as for the logistic regression approach just described, but for biomarker level given case status and covariates rather than just covariates. The scale parameter is assumed to be constant within cases and within controls, and the shape parameter log-linear in case status and covariates. This implies that the log-odds ratio of interest varies with the biomarker level and covariate values. However, in the special case where the shape-parameter coefficient for case status is 0, the log-odds ratio reduces to the difference in the inverse scale parameters for controls and cases. Thus, one can test whether the relevant coefficient is 0, and then either report the estimated log-odds ratio and standard error, or use graphical methods to visualize the non-constant effect. This is analogous to the normal discriminant function approach, where different residual error variances for cases and controls corresponds to a second-order effect on log-odds of disease (Cornfield *et al.*, 1962; Lyles *et al.*, 2009).

The Gamma discriminant function setup lends itself to pooled data because of the pool-wise sum result for Gamma variates. If the model holds for individual-level biomarker levels, then a very similar model applies to the summed biomarker level. The setup also permits in-

corporating errors; we pursue the same multiplicative lognormal error structure as proposed for the Gamma logistic regression.

Our motivating example is the same as in Chapter 2: estimation of the covariate-adjusted odds ratio relating levels of a cytokine to odds of spontaneous abortion (SA). We use AIC to confirm better model fit for the Gamma analogues of the Chapter 2 methods and perform simulations to confirm validity, assess the degree to which AIC can identify the correct model, and gauge whether estimation is reasonably stable in scenarios where the Gamma models are uniquely identifiable.

3.2 Methods

3.2.1 Scenario

The goal is to estimate the log-OR relating Y to X adjusted for $\boldsymbol{C}$ using poolwise data. As in Chapter 2, we consider a design in which the i^{th} pool is comprised of g_i cases or controls. Let $Y_i = 1$ for case pools and 0 for control pools. The assay is assumed to target the mean biomarker level for members of each pool, $\bar{X}_i = \frac{1}{g_i}\sum_{j=1}^{g_i} X_{ij}$, from which the sum can be calculated as $X_i^* = g_i\bar{X}_i$. We assume individual-level covariates are available, say $\boldsymbol{C}_i = (\boldsymbol{C}_{i1}, ..., \boldsymbol{C}_{ig_i})^T$, as well as summed covariates $\boldsymbol{C}_i^* = \sum_{j=1}^{g_i} \boldsymbol{C}_{ij}$.

Rather than observing the precise $\bar{X}_i$, we assume the assay produces an error-contaminated version $\tilde{\bar{X}}_i$ affected by processing error and/or measurement error. To allow for replicates, suppose there are $k_i \geq 1$ such measurements, such that we observe $\tilde{\tilde{\boldsymbol{X}}}_i = (\tilde{\tilde{X}}_{i1}, ..., \tilde{\tilde{X}}_{ik_i})^T$ and can calculate the sums $\tilde{\boldsymbol{X}}_i^* = (\tilde{X}_{i1}^*, ..., \tilde{X}_{ik_i}^*)^T$. In this case, replicates are not necessarily single-specimen pools; if $g_i > 1$, we assume the k_i measurements are affected by the same processing error ϵ_i^p.

3.2.2 Logistic regression methods

Homogeneous pools logistic regression

Absent errors, Weinberg and Umbach (1999) showed that the appropriate logistic regression model for analyzing poolwise data based on observed $(Y_i, X_i^*, \boldsymbol{C}_i^*)$ values is:

$$\text{logit}[P(Y_i = 1)] = q_i + g_i\beta_0 + \beta_x X_i^* + \boldsymbol{\beta}_{\boldsymbol{c}}^T \boldsymbol{C}_i^* \tag{3.49}$$

where the offset is given by:

$$q_i = g_i \text{log}(\frac{P(A|D)}{P(A|\bar{D})}) + g_i \text{log}(\frac{n_{\bar{D}}}{n_D}) + \text{log}(\frac{\#\text{ case pools of size } g_i}{\#\text{ control pools of size } g_i}) \tag{3.50}$$

Further details on the offset were provided previously (see Section 2.2.1, pg. 40).

General likelihood setup

Regardless of the form of $X|\boldsymbol{C}$ and whether errors are assumed to be additive or multiplicative, the likelihood contribution for the i^{th} pool given the observed $(Y_i, \tilde{\boldsymbol{X}}_i^*, \boldsymbol{C}_i)$ is $L_i(\boldsymbol{\theta}) \propto f(Y_i, \tilde{\boldsymbol{X}}_i^* | \boldsymbol{C}_i)$. This can be factored as:

$$\begin{aligned} L_i(\boldsymbol{\theta}) &= \int_{X_i^*} f(Y_i | \tilde{\boldsymbol{X}}_i^*, X_i^*, \boldsymbol{C}_i)\ f(\tilde{\boldsymbol{X}}_i^* | X_i^*, \boldsymbol{C}_i)\ f(X_i^* | \boldsymbol{C}_i)\ dX_i^* \\ &= \int_{X_i^*} f(Y_i | X_i^*, \boldsymbol{C}_i^*)\ f(\tilde{\boldsymbol{X}}_i^* | X_i^*)\ f(X_i^* | \boldsymbol{C}_i)\ dX_i^* \end{aligned} \tag{3.51}$$

The first term simplifies as above based on two results: in the poolwise logistic regression model Eq. 3.49, individual-level $\boldsymbol{C}_{ij}$'s are not needed, only the summed $\boldsymbol{C}_i^*$'s; and we make a standard assumption that the imprecise $\tilde{\boldsymbol{X}}_i^*$ does not additionally inform Y given the true X_i^* and $\boldsymbol{C}_i^*$. For the second term, we assume that errors are unrelated to covariate values.

In Eq. 3.51, the first density is specified by Eq. 3.49, while the second and third require additional assumptions for the errors and biomarker.

Additive errors, normal biomarker

This approach, which we term "NLR" for normal logistic regression here, is described in the previous chapter. Briefly, we assume the error structure:

$$\tilde{\bar{\boldsymbol{X}}}_i = \mathbf{1}_{k_i}\bar{X}_i + \mathbf{1}_{k_i}\epsilon_i^p I(g_i > 1) + \boldsymbol{\epsilon}_i^m \tag{3.52}$$

With normality and independence assumptions on the errors, the second term in Eq. 3.51 is given by:

$$\tilde{\boldsymbol{X}}_i^* | X_i^* \sim N_{k_i}\left(\mathbf{1}_{k_i}X_i^*,\ g_i^2\sigma_p^2 I(g_i > 1)\boldsymbol{J}_{k_i} + g_i^2\sigma_m^2\boldsymbol{I}_{k_i}\right) \tag{3.53}$$

For the third term, a normal linear regression for the individual data leads to:

$$X_i^* | \boldsymbol{C}_i \sim N(g_i\alpha_0 + \boldsymbol{\alpha}_{\boldsymbol{C}}^T\boldsymbol{C}_i^*,\ g_i\sigma_x^2) \tag{3.54}$$

Multiplicative errors, Gamma biomarker

Next we propose a Gamma $X|\boldsymbol{C}$ logistic regression approach ("GLR"), motivated by the possibility of skewed, strictly positive biomarkers, for which the NLR assumptions may not be justified. The likelihood Eq. 3.51 still applies, but we use different models for the second and third densities. Rather than mean-0 normal errors acting additively on the poolwise means, we assume mean-1 lognormal errors acting multiplicatively (Carroll *et al.*, 2006). The analogue of Eq. 3.52 is:

$$\tilde{\bar{\boldsymbol{X}}}_i = \boldsymbol{\epsilon}_i^m(\epsilon_i^p)^{I(g_i>1)}\bar{X}_i \tag{3.55}$$

with the error assumptions $\boldsymbol{\epsilon}_i^m \overset{ind}{\sim} LN_{k_i}(-\frac{\sigma_m^2}{2}\mathbf{1}_{k_i}, \sigma_m^2\boldsymbol{I}_{k_i})$, $\epsilon_i^p \overset{iid}{\sim} LN(-\frac{\sigma_p^2}{2}, \sigma_p^2)$, and $\boldsymbol{\epsilon}_i^m \perp\!\!\!\perp \epsilon_i^p$.

To determine the form of $\tilde{\boldsymbol{X}}_i^* | X_i^*$, first note that $\tilde{\boldsymbol{X}}_i^* = g_i\tilde{\bar{\boldsymbol{X}}}_i = \boldsymbol{\epsilon}_i^m(\epsilon_i^p)^{I(g_i>1)}X_i^*$. The prod-

uct of the independent lognormal error terms, $\boldsymbol{\epsilon}_i = \boldsymbol{\epsilon}_i^m (\epsilon_i^p)^{I(g_i>1)}$, is multivariate lognormal:

$$\boldsymbol{\epsilon}_i \sim LN_{k_i}\left(\mathbf{1}_{k_i}\left[-\frac{1}{2}(\sigma_p^2 I(g_i>1)+\sigma_m^2)\right],\ \sigma_p^2 I(g_i>1)\boldsymbol{J}_{k_i}+\sigma_m^2\boldsymbol{I}_{k_i}\right) \tag{3.56}$$

So $\tilde{\boldsymbol{X}}_i^*|X_i^*$ is also multivariate lognormal:

$$\tilde{\boldsymbol{X}}_i^*|X_i^* \sim LN_{k_i}\left(\mathbf{1}_{k_i}\left[\log(X_i^*)-\frac{1}{2}(\sigma_p^2 I(g_i>1)+\sigma_m^2)\right],\ \sigma_p^2 I(g_i>1)\boldsymbol{J}_{k_i}+\sigma_m^2\boldsymbol{I}_{k_i}\right) \tag{3.57}$$

which can be viewed as a multiplicative lognormal analogue of Eq. 3.53.

For $X|\boldsymbol{C}$, we assume a constant-scale Gamma model:

$$X_{ij}|\boldsymbol{C}_{ij} \sim \text{Gamma}\left(\alpha_{ij}=e^{\alpha_0+\boldsymbol{\alpha}_{\boldsymbol{c}}^T\boldsymbol{C}_{ij}},\ \beta_{ij}=b\right) \tag{3.58}$$

This implies $E(X_{ij}|\boldsymbol{C}_{ij}) = \alpha_{ij}\beta_{ij} = be^{\alpha_0+\boldsymbol{\alpha}_{\boldsymbol{c}}^T\boldsymbol{C}_{ij}}$, or that there is a monotone, nonlinear relationship between each covariate and the expected value of the biomarker. It also implies $V(X_{ij}|\boldsymbol{C}_{ij}) = \alpha_{ij}\beta_{ij}^2 = b^2e^{\alpha_0+\boldsymbol{\alpha}_{\boldsymbol{c}}^T\boldsymbol{C}_{ij}}$, which means $V(X_{ij}) = bE(X_{ij}|\boldsymbol{C}_{ij})$, or the variance is directly proportional to the mean.

The sum of independent Gamma variables with shape parameter α_j and the same scale parameter β is Gamma$(\sum_j \alpha_j, \beta)$, so the poolwise sum biomarker level is distributed:

$$X_i^*|\boldsymbol{C}_i \sim \text{Gamma}\left(\alpha_i=\sum_{j=1}^{g_i} e^{\alpha_0+\boldsymbol{\alpha}_{\boldsymbol{c}}^T\boldsymbol{C}_{ij}},\ \beta_i=b\right) \tag{3.59}$$

and the likelihood Eq. 3.51 is completely specified.

3.2.3 Discriminant function methods

General likelihood setup

For a discriminant function analysis, the log-OR of interest is targeted via a model for the continuous variable given the outcome and covariates, $X|(Y, \boldsymbol{C})$. The likelihood contribution for the i^{th} pool given the observed $(Y_i, \tilde{\boldsymbol{X}}_i^*, \boldsymbol{C}_i)$ is $L_i(\boldsymbol{\theta}) \propto f(\tilde{\boldsymbol{X}}_i^*|Y_i, \boldsymbol{C}_i)$, which can be factored:

$$\begin{aligned} L_i(\boldsymbol{\theta}) &= \int_{X_i^*} f(\tilde{\boldsymbol{X}}_i^*, X_i^*|Y_i, \boldsymbol{C}_i)\ dX_i^* \\ &= \int_{X_i^*} f(\tilde{\boldsymbol{X}}_i^*|X_i^*)\ f(X_i^*|Y_i, \boldsymbol{C}_i)\ dX_i^* \end{aligned} \tag{3.60}$$

Additive errors, normal biomarker

The normal discriminant function approach ("NDFA") was described in Chapter 2. Briefly, $\tilde{\boldsymbol{X}}_i|X_i^*$ is the same as for NLR (Eq. 3.53), and for $X_i^*|(Y_i, \boldsymbol{C}_i)$ we assume:

$$X_i^*|(Y_i, \boldsymbol{C}_i) \sim N\left(g_i\gamma_0 + \gamma_y(g_iY_i) + \boldsymbol{\gamma_c}^T\boldsymbol{C}_i^*,\ g_i\sigma_{y_i}^2\right) \tag{3.61}$$

Note that we use g_iY_i here rather than Y_i^* for ease of notation, since we are assuming homogeneous pools. The residual error variance $\sigma_{y_i}^2$ is σ_1^2 for cases and σ_0^2 for controls.

Previously we assumed $\sigma_1^2 = \sigma_0^2 = \sigma^2$, which leads to a constant log-OR scenario, but here we consider the possibility of different error variances for cases and controls. Applying Bayes rule to the individual-level model, with subscripts omitted for simplicity, $P(Y = 1|X, \boldsymbol{C}) = \frac{f(X|Y=1,\boldsymbol{C})P(Y=1|\boldsymbol{C})}{f(X|\boldsymbol{C})}$ and $P(Y = 0|X, \boldsymbol{C}) = \frac{f(X|Y=0,\boldsymbol{C})P(Y=0|\boldsymbol{C})}{f(X|\boldsymbol{C})}$. Thus:

$$\begin{aligned} \text{logit}[P(Y = 1|X, \boldsymbol{C})] &= \text{logit}[P(Y = 1|\boldsymbol{C})] + \log[f(X|Y = 1, \boldsymbol{C})] - \log[f(X|Y = 0, \boldsymbol{C})] \\ &= d + X^2\left(\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2}\right) + X\left(\frac{\gamma_y}{\sigma_1^2} + (\gamma_0 + \boldsymbol{\gamma_c}^T\boldsymbol{C})(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2})\right) \end{aligned} \tag{3.62}$$

where d is logit$[P(Y=1|\boldsymbol{C})]$ plus constant terms involving $(\boldsymbol{\gamma}^T, \sigma_0^2, \sigma_1^2)$. For a 1-unit increase in X, the log-OR is:

$$\text{log-OR} = \frac{\gamma_y}{\sigma_1^2} + \left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right)\left(X - \gamma_0 - \boldsymbol{\gamma_c}^T\boldsymbol{C} + \frac{1}{2}\right) \tag{3.63}$$

which simplifies to $\frac{\gamma_y}{\sigma^2}$ when $\sigma_1^2 = \sigma_0^2 = \sigma^2$. For $\sigma_1^2 \neq \sigma_0^2$, the log-OR varies with X as well as $\boldsymbol{C}$. In that case, one could visualize the association by plotting the log-OR vs. X over the range of X in the data, with $\boldsymbol{C}$ held fixed at certain covariate values (e.g. means), or for several sets of covariate values. Confidence bands based on the delta method could also be included. For $\widehat{\text{log-OR}} = f(\hat{\boldsymbol{\theta}}) = f(\hat{\boldsymbol{\gamma}}, \hat{\sigma}_1^2, \hat{\sigma}_0^2)$, the Jacobian derivatives are:

$$\begin{aligned}
\frac{\partial f(\boldsymbol{\theta})}{\partial \gamma_0} &= \frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2} \\
\frac{\partial f(\boldsymbol{\theta})}{\partial \gamma_y} &= \frac{1}{\sigma_1^2} \\
\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma_c}^T} &= \left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right)\boldsymbol{C}^T \\
\frac{\partial f(\boldsymbol{\theta})}{\partial \sigma_1^2} &= \frac{1}{\sigma_1^4}\left(X - \gamma_0 - \gamma_y - \boldsymbol{\gamma_c}^T\boldsymbol{C} + \frac{1}{2}\right) \\
\frac{\partial f(\boldsymbol{\theta})}{\partial \sigma_0^2} &= -\frac{1}{\sigma_0^4}\left(X - \gamma_0 - \boldsymbol{\gamma_c}^T\boldsymbol{C} + \frac{1}{2}\right)
\end{aligned} \tag{3.64}$$

Multiplicative errors, Gamma biomarker

For a Gamma discriminant function approach ("GDFA"), we use the same multiplicative error assumptions as for GLR, such that $\tilde{\boldsymbol{X}}_i|X_i^*$ is given by Eq. 3.57. For $X_i^*|(Y_i, \boldsymbol{C}_i)$, we assume the following individual-level model:

$$X_{ij}|(Y_{ij}, \boldsymbol{C}_{ij}) \sim \text{Gamma}\left(\alpha_{ij} = e^{\gamma_0 + \gamma_y Y_{ij} + \boldsymbol{\gamma_c}^T\boldsymbol{C}_{ij}},\ \beta_{ij} = b_{y_{ij}}\right) \tag{3.65}$$

If the i^{th} pool is comprised of g_i cases or controls, then a similar Gamma model applies to $X_i^*|(Y_i, \boldsymbol{C}_i)$:

$$X_i^*|(Y_i, \boldsymbol{C}_i) \sim \text{Gamma}\left(\alpha_i = \sum_{j=1}^{g_i} e^{\gamma_0 + \gamma_y Y_i + \boldsymbol{\gamma_c}^T \boldsymbol{C}_i},\ \beta_i = b_{y_i}\right) \tag{3.66}$$

This model is compatible with data on hand, which is assumed to include individual-level covariates. Note that the above result requires that members of a pool have the same scale parameter, which means pools have to be homogeneous with respect to case status. This was not a requirement for NDFA.

For the log-OR implied by this model, applying Bayes rule and taking the logit leads to:

$$\text{logit}[P(Y=1|X, \boldsymbol{C})] = d + \log(x)\left(e^{\gamma_0 + \gamma_y + \boldsymbol{\gamma_c}^T \boldsymbol{C}} - e^{\gamma_0 + \boldsymbol{\gamma_c}^T \boldsymbol{C}}\right) + x\left(\frac{1}{b_0} - \frac{1}{b_1}\right) \tag{3.67}$$

The log-OR for a 1-unit increase in X is:

$$\text{log-OR} = \frac{1}{b_0} - \frac{1}{b_1} + \log\left(\frac{X+1}{X}\right) e^{\gamma_0 + \boldsymbol{\gamma_c}^T \boldsymbol{C}}(e^{\gamma_y} - 1) \tag{3.68}$$

which simplifies to $\frac{1}{b_0} - \frac{1}{b_1}$ when $\gamma_y = 0$. For $\gamma_y \neq 0$, the same approach could be taken as for NDFA: plot the log-OR vs. X over the range of X in the data, at fixed $\boldsymbol{C}$ values, and including confidence bands. The Jacobian derivatives are:

$$\begin{aligned}
\frac{\partial f(\boldsymbol{\theta})}{\partial \gamma_0} &= \log\left(\frac{X+1}{X}\right) e^{\gamma_0 + \boldsymbol{\gamma_c}^T \boldsymbol{C}}(e^{\gamma_y} - 1) \\
\frac{\partial f(\boldsymbol{\theta})}{\partial \gamma_y} &= \log\left(\frac{X+1}{X}\right) e^{\gamma_0 + \gamma_y + \boldsymbol{\gamma_c}^T \boldsymbol{C}} \\
\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma_c}^T} &= \log\left(\frac{X+1}{X}\right) e^{\gamma_0 + \boldsymbol{\gamma_c}^T \boldsymbol{C}}(e^{\gamma_y} - 1)\boldsymbol{C}^T \\
\frac{\partial f(\boldsymbol{\theta})}{\partial b_1} &= \frac{1}{b_1^2} \\
\frac{\partial f(\boldsymbol{\theta})}{\partial b_0} &= -\frac{1}{b_0^2}
\end{aligned} \tag{3.69}$$

3.2.4 Implementation

We use similar computational methods for all four corrective methods. NDFA is unique in that it has a closed-form likelihood; the others require numerical integration. For NLR, we use the full ML approach rather than approximate ML. We use the R function *hcubature* in **cubature** v. 1.3-11 (Narasimhan and Johnson, 2017) for numerical integration, the function *hessian* in **pracma** v. 2.1.1 (Borchers, 2017) for approximating Hessian matrices, and the function *nlminb* in base R for maximizing likelihoods.

We added several functions to our R package **pooling** (Van Domelen, 2018b) to implement the methods in this chapter. The functions *p_logreg_xerrors* and *p_logreg_xerrors2* are for NLR and GLR, and the functions *p_dfa_xerrors* and *p_dfa_xerrors2* are for NDFA and GDFA, respectively.

For the discriminant function methods, there is an option for whether to assume a constant log-OR, e.g. set $\sigma_1^2 = \sigma_0^2$ for NDFA and $\gamma_y = 0$ for GDFA. There is also an option to perform a likelihood ratio test to formally test these hypotheses. Additionally, when the log-OR is not assumed constant, the functions *plot_dfa* and *plot_dfa2* can be used to graph the estimated log-OR vs. biomarker level at fixed covariate values.

3.3 Results

3.3.1 Motivating example

We used the same dataset as in Chapter 2, from the Collaborative Perinatal Project (CPP), to explore whether the cytokine monocyte chemotactic protein (*MCP-1*) is associated with risk of spontaneous abortion (SA) controlling for mother's age, race, and current smoking.

Given that the 126 single-specimen pools are not subject to processing error, and the 30 replicates suggest a relatively small amount of measurement error (Figure 3.7, left), a histogram of the singles should give a reasonable indication of the marginal *MCP-1* distribution (Figure 3.7, right). The data are more compatible with lognormal and Gamma distributions than normal.

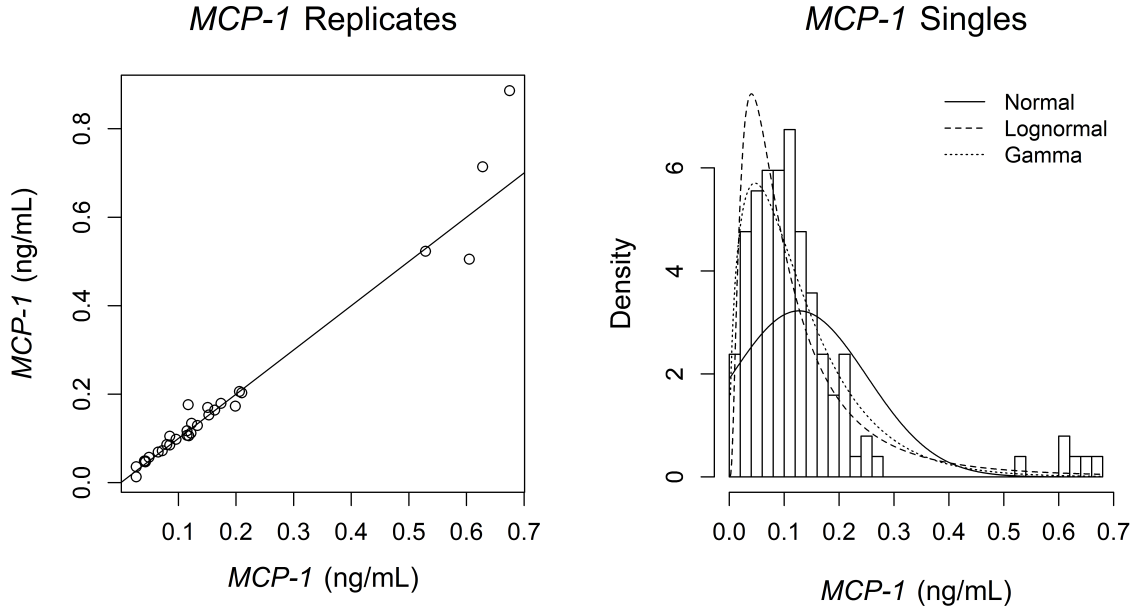

Figure 3.7: Agreement between two *MCP-1* measurements for 30 single-specimen pools (left) and histogram of all 126 singles (right) in CPP.

Table 3.13 summarizes model fits for the two logistic regression models, using all available data including replicates and modeling both error types. Covariates C_1-C_3 represent mother's age, non-white race, and current smoking; X represents the pooled exposure *MCP-1*, multiplied by 10 so that the log-OR is for a 0.1-ng/mL increment. The β's are logistic regression coefficients in Eq. 3.49, α's are coefficients in the $X|\boldsymbol{C}$ models (i.e. Eq. 3.54 and Eq. 3.59), b is the scale parameter in the Gamma $X|\boldsymbol{C}$ model (Eq. 3.59), σ_x^2 is the residual error variance in the normal $X|\boldsymbol{C}$ model (Eq. 3.54), and (σ_p^2, σ_m^2) are the processing and measurement error variances, respectively.

Table 3.13: Logistic regression estimates for odds of spontaneous abortion in CPP. Values are point estimates (SE).

		(AIC = 2340.8)	(AIC = 1787.5)
	Naive	Normal logistic regression	Gamma logistic regression
β_0	-1.57 (0.37)	-1.58 (0.37)	-1.60 (0.39)
β_x	0.01 (0.02)	0.05 (0.08)	0.05 (0.12)
β_{c_1}	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)
β_{c_2}	0.56 (0.18)	0.57 (0.18)	0.57 (0.18)
β_{c_3}	0.34 (0.16)	0.34 (0.16)	0.34 (0.16)
α_0	-	0.50 (0.38)	0.38 (0.26)
α_{c_1}	-	0.03 (0.01)	0.01 (0.01)
α_{c_2}	-	-0.17 (0.17)	-0.33 (0.11)
α_{c_3}	-	0.02 (0.16)	-0.01 (0.09)
b	-	-	0.69 (0.08)
σ_x^2	-	1.58 (0.21)	-
σ_p^2	-	0.73 (0.18)	0.62 (0.09)
σ_m^2	-	0.11 (0.03)	0.02 (0.01)

AIC favored GLR over NLR. The estimated log-OR was higher for NLR and GLR than for the naive poolwise logistic regression fit ignoring errors, but still not significantly different than 0. The other logistic regression coefficients were virtually identical for the three methods. Both NLR and GLR suggested much more severe processing errors than measurement errors.

Notably, if the replicate *MCP-1* measurements had not been included, the NLR model could not be fit with both error types, while the GLR model could. Identifiability for NLR requires at least three different pool sizes including 1; the CPP data only has pools of size 1 and 2. GLR fit without replicates gave a somewhat larger log-OR estimate ($\hat{\beta}_x = 0.08,\ SE = 0.12$) and very different variance estimates ($\hat{\sigma}_p^2 = 0.17, \hat{\sigma}_m^2 = 0.42$) compared to the fit with replicates. It is unclear whether GLR's identifiability is practical in this scenario, given the different (σ_p^2, σ_m^2) estimates and NLR's instability without replicates in Chapter 2 (Figure

Table 3.14: Discriminant function estimates for odds of spontaneous abortion in CPP. Values are point estimates (SE).

	(AIC = 1796.5)	(AIC = 1242.9)
	Normal discriminant function	Gamma discriminant function
γ_0	0.50 (0.38)	0.41 (0.26)
γ_y	0.08 (0.13)	-
γ_{c_1}	0.02 (0.01)	0.01 (0.01)
γ_{c_2}	-0.19 (0.17)	-0.34 (0.11)
β_{c_3}	0.01 (0.16)	-0.02 (0.09)
b_1	-	0.72 (0.09)
b_0	-	0.67 (0.09)
σ^2	1.58 (0.21)	-
σ_p^2	0.73 (0.18)	0.62 (0.09)
σ_m^2	0.11 (0.03)	0.02 (0.01)
log-OR	0.05 (0.08)	0.10 (0.13)

2.5), in what seemed to be an easier identifiability scenario with three pool sizes. It may be that GLR's unique identifiability is not practical when the pooled biomarker is close to being normally distributed (Carroll *et al.*, 2006). We explore this issue later via simulations.

Table 3.14 summarizes fits for the two discriminant function methods, under the assumption that the log-OR is constant with X. Results were similar to Table 3.13 in that AIC values favored the Gamma approach over normal, and neither suggested a significant association for *MCP-1*. Also mirroring the logistic regression results, GDFA parameters were identifiable without replicates ($\widehat{\text{log-OR}} = 0.08,\ SE = 0.11,\ \hat{\sigma}_p^2 = 0.12,\ \hat{\sigma}_m^2 = 0.44$), while NDFA parameters were not.

The constant log-OR models reported in Table 3.14 are the result of restrictions corresponding to testable hypotheses. For NDFA, the log-OR is constant if the residual error variance in the $X|(Y, \boldsymbol{C})$ model is the same for cases and controls, i.e. under $H_0 : \sigma_1^2 = \sigma_0^2$. For GDFA, it is constant under $H_0^* : \gamma_y = 0$. Likelihood ratio tests did not reject H_0

($D = 0.85,\ p = 0.36$) or H_0^* ($D = 0.91,\ p = 0.34$) in models fit with replicates and both error types.

Despite little evidence to favor the non-constant log-OR models, we plotted the estimated log-OR vs. *MCP-1* based on these fitted models for the four combinations of the two binary covariates and mother's age held fixed at 26 years. These plots are shown in Figure 3.8, with the x-axis corresponding to the range of *MCP-1* values for the 126 singles (0.02 to 6.75, median = 1.025). The graphs suggest a harmful effect at low *MCP-1* levels and a protective effect at higher levels; however, the confidence bands are compatible with no association over the entire range.

In summary, the Gamma models fit the CPP data better than the corresponding normal models, and were unique in that they could be fit without replicates. Substantive results were similar for all four methods: the estimated log-OR is small, there is little evidence of an association between *MCP-1* and risk of SA, and poolwise *MCP-1* measurements seem to be more severely impacted by processing error than by measurement error.

3.3.2 Simulations

The purpose of the first simulation study is to confirm validity of the Gamma methods and assess robustness of the four methods under model misspecification. Data were generated under either GLR or GDFA, mimicking the CPP data and estimated parameters, and the log-OR estimated based on the data-generating model as well as the three others.

For each trial under GLR, individual-level covariates (C_1 = mother's age, C_2 = non-white race, C_3 = current smoking) were generated independently for 686 subjects as follows: $C_1 \in (14, ..., 45)$ with sampling probabilities equal to the CPP proportions; $C_2 \sim$ Bernoulli(0.34); and $C_3 \sim$ Bernoulli(0.47). Individual-level X (*MCP-1*) were then generated from Eq. 3.58 with $\alpha_0 = 0.38$, $\boldsymbol{\alpha_c} = (0.01, -0.33, -0.01)^T$, $b = 0.69$, and Y (SA) generated from Eq.

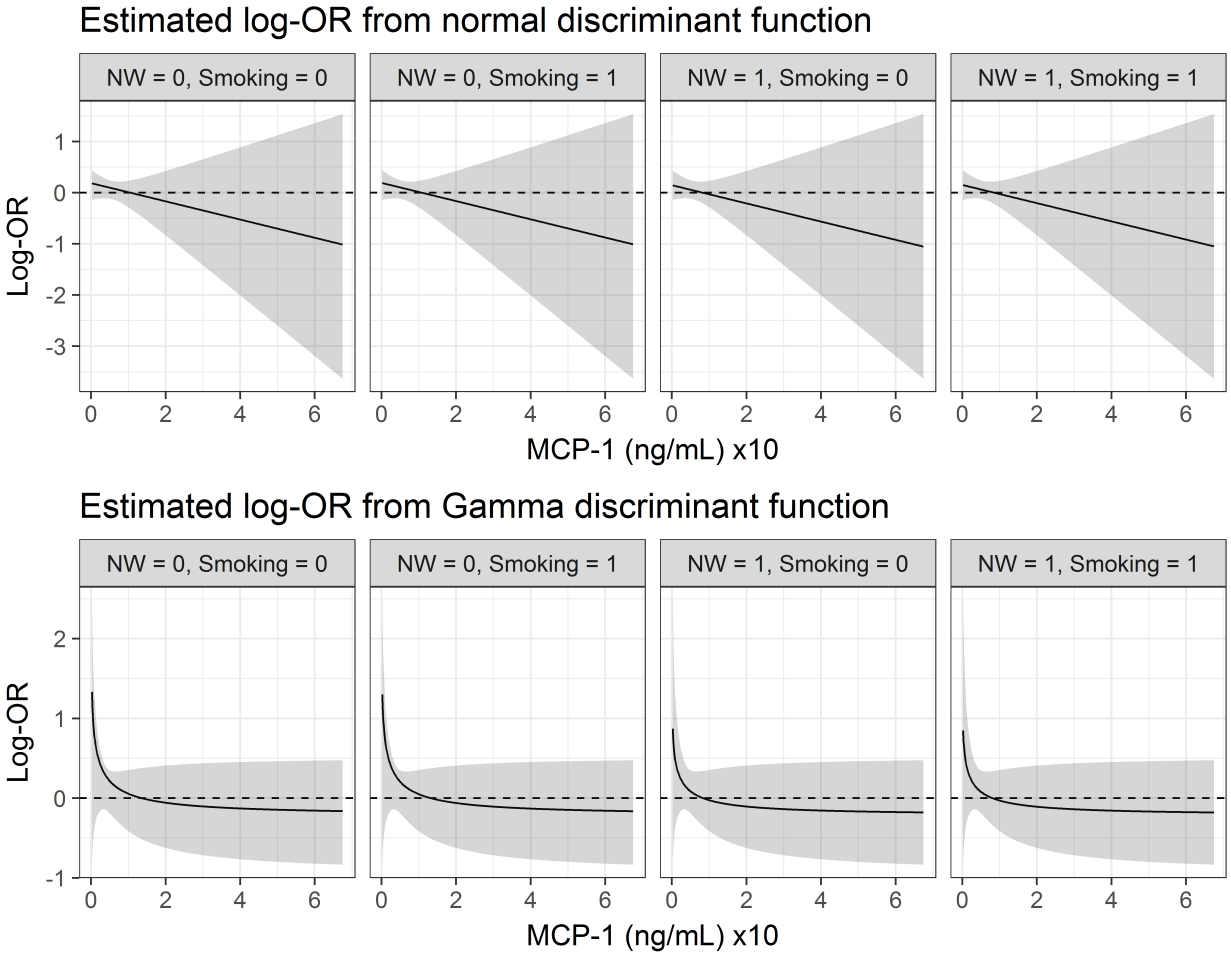

Figure 3.8: Estimated log-OR (95% confidence band) vs. *MCP-1* from fitted normal (top) and Gamma (bottom) discriminant function models in CPP.

2.27 with $\beta_0 = -1.60$, $\beta_x = 0.15$ (increased from $\hat{\beta}_x = 0.05$), $\boldsymbol{\beta_c} = (0.04, 0.57, 0.34)^T$. Observations were then split into n_1 cases and n_0 controls. The n_1 cases were randomly formed into $\frac{n_1}{3}$ (rounded up) pools of size 2 and the rest left as singles, and similarly for the n_0 controls. Poolwise means $\bar{X}_i$ were then calculated, multiplied by lognormal processing errors with $\sigma_p^2 = 0.62$ (if $g_i = 2$) and lognormal measurement errors with $\sigma_m^2 = 0.02$, and multiplied by g_i to produce imprecise poolwise sums $\tilde{X}_i^*$. For 30 randomly selected pools, $\tilde{\boldsymbol{X}}_i = (\tilde{X}_{i1}, \tilde{X}_{i2})^T$ was generated from the same process but for two independent measurement errors rather than one.

For GDFA, $\boldsymbol{C}$ was generated via the same process, $Y|\boldsymbol{C}$ based on a logistic regression with $\beta_0^* = -1.64$ and $\boldsymbol{\beta_c^*} = (0.04, 0.57, 0.36)^T$, and $X|(Y, \boldsymbol{C})$ based on Eq. 3.65 with $\gamma_0 = 0.41$, $\gamma_y = 0$, $\boldsymbol{\gamma_c} = (0.01, -0.34, -0.02)^T$, $b_1 = 0.7449$ (increased from $\hat{b}_1 = 0.72$ to induce a log-OR of 0.15), and $b_0 = 0.67$. Poolwise data was generated via the same process as above, again with $\sigma_p^2 = 0.62$ and $\sigma_m^2 = 0.02$.

Results are summarized in Table 3.15. For data generated under GLR, the naive poolwise logistic regression (i.e. ignoring processing error and measurement error) underestimated the true log-OR and had poor CI coverage. The correctly specified GLR estimator appeared virtually unbiased with nominal coverage; GDFA performed about the same as GLR. NLR and NDFA performed surprisingly well despite assuming additive normal rather than multiplicative lognormal errors; they were unbiased, only slightly less efficient than the Gamma methods, and had close to nominal coverage. These trends were nearly identical for data generated under GDFA. For data generated under GLR, AIC favored GLR over NLR in every trial; for data generated under GDFA, AIC favored GDFA over NDFA in every trial.

The next set of simulations is aimed at gauging whether the Gamma models' unique identifiability absent replicates is practically useful. We consider the CPP scenario: pools of size 1 and 2 and poolwise biomarker measurements subject to multiplicative lognormal processing errors and measurement errors. Data were generated under GLR and GDFA in

Table 3.15: Simulation results for estimation of adjusted log-OR for *MCP-1* and spontaneous abortion (500 trials each, true log-OR = 0.15).

	Mean bias	SD	Mean SE	MSE	Coverage
Data generated GLR[a]					
Naive	-0.104	0.053	0.050	0.014	0.443
GLR	0.006	0.125	0.127	0.016	0.966
NLR	0.003	0.132	0.134	0.017	0.966
GDFA	0.002	0.123	0.125	0.015	0.962
NDFA	0.001	0.128	0.132	0.016	0.966
Data generated GDFA					
Naive	-0.105	0.053	0.049	0.014	0.388
GLR	0.002	0.124	0.124	0.015	0.962
NLR	0.001	0.137	0.131	0.019	0.944
GDFA	0.000	0.121	0.121	0.015	0.956
NDFA	0.000	0.135	0.130	0.018	0.942

[a] Excludes 3 trials in which variance-covariance matrix for NLR was not positive definite.

Table 3.16: Simulation results for estimation of log-OR with pools of size 1 and 2, with and without replicates (500 trials each, true log-OR = 0.15).

	Gamma logistic regression				*Gamma discriminant function*			
	Mean bias	Median bias	CI coverage	Median CI width	Mean bias	Median bias	CI coverage	Median CI width
n = 686								
No replicates	0.006[a]	-0.007[a]	0.976[a]	0.573[a]	1.113[c]	0.033[c]	0.942[c]	0.553[c]
30 replicates	-0.004[a]	-0.009[a]	0.970[a]	0.547[a]	0.021[c]	0.013[c]	0.954[c]	0.521[c]
n = 2,000								
No replicates	0.003[b]	0.004[b]	0.952[b]	0.323[b]	0.003[d]	0.004[d]	0.958[d]	0.307[d]
30 replicates	0.001[b]	0.001[b]	0.950[b]	0.318[b]	0.002[d]	0.001[d]	0.954[d]	0.303[d]

[a] Excludes 2 trials with non-positive definite variance-covariance matrix.

[b] Excludes 1 trial with non-positive definite variance-covariance matrix.

[c] Excludes 2 trials with non-positive definite variance-covariance matrix.

[d] Excludes 5 trials with non-positive definite variance-covariance matrix.

the same manner as in previous simulations, but for various sample sizes, with and without the 30 replicates. After initially observing good performance without replicates despite $\hat{\sigma}_m^2$ frequently hitting the lower bound of 0.001, simply because the measurement error was small enough to ignore, we increased σ_m^2 from 0.02 to 0.2 and decreased σ_p^2 from 0.62 to 0.42. In trials where $\hat{\sigma}_p^2$ or $\hat{\sigma}_m^2$ hit 0.001, processing error-only and measurement error-only models were fit, and the one with the lower AIC selected. Results are summarized in Table 3.16.

Overall performance was surprisingly good for the no-replicates estimators, perhaps with the exception of the $n = 686$ GDFA scenario, where there was mean bias due to 8 trials where $\widehat{\log\text{-OR}}$ was outside of $(-1, 1)$. Confidence intervals were wider without replicates, but not much, especially for $n = 2,000$. The measurement error variance estimate $\hat{\sigma}_m^2$ occasionally hit 0.001 for the no replicates scenarios (2.8% of trials for GLR and $n = 686$, 4.3% of trials for GDFA and $n = 686$, and 0.3% of trials for GDFA and $n = 2,000$).

Lastly, we compare efficiency of a pooling vs. traditional design for the same number of total assays in a no measurement error (processing error only) scenario, where the pooling

design is perhaps most appealing. We generated individual-level data under the Gamma discriminant function model with $n = 686$ and split the data into cases and controls. For the pooling design, we formed $\frac{n_1}{4.5}$ (rounded up) case pools of size 4 and left the remaining cases as singles, and similarly for controls, to produce approximately 2x as many pools of size 4 as pools of size 1. For the traditional design, we randomly sampled the same number of cases and controls as there were case pools and control pools in the same trial and obtained individual-level X values, which were precise since singles are not affected by processing error. Figure 3.9 shows that the pooling design was more efficient for small σ_p^2, but that efficiency advantage eroded and eventually reversed as σ_p^2 was increased. This is consistent with Chapter 2 efficiency results under additive normal processing errors (see top panel of Figure 2.6).

3.4 Discussion

We have developed two Gamma model-based methods for estimating the adjusted log-odds ratio relating a binary outcome to a continuous exposure measured in pools and subject to errors. This work integrates the poolwise logistic regression approach of Weinberg and Umbach (1999) with the error modeling assumptions of Schisterman *et al.* (2010) and the discriminant function ideas of Lyles *et al.* (2015) and Whitcomb *et al.* (2012). Accommodating skewed, positive biomarkers should broaden the scope of scenarios where a highly cost-effective homogeneous pools study design can be utilized.

First, we wish to emphasize the utility of our methods, which could appear narrowly focused (homogeneous pools design + skewed biomarker + errors). The homogeneous pools design is compelling as it offers potentially large gains in statistical power over a traditional design (e.g. Figure 3.9 with $\sigma_p^2 = 0$). Absent errors, there would be no need to worry about the distribution of the pooled biomarker; one could simply fit the Weinberg and Umbach

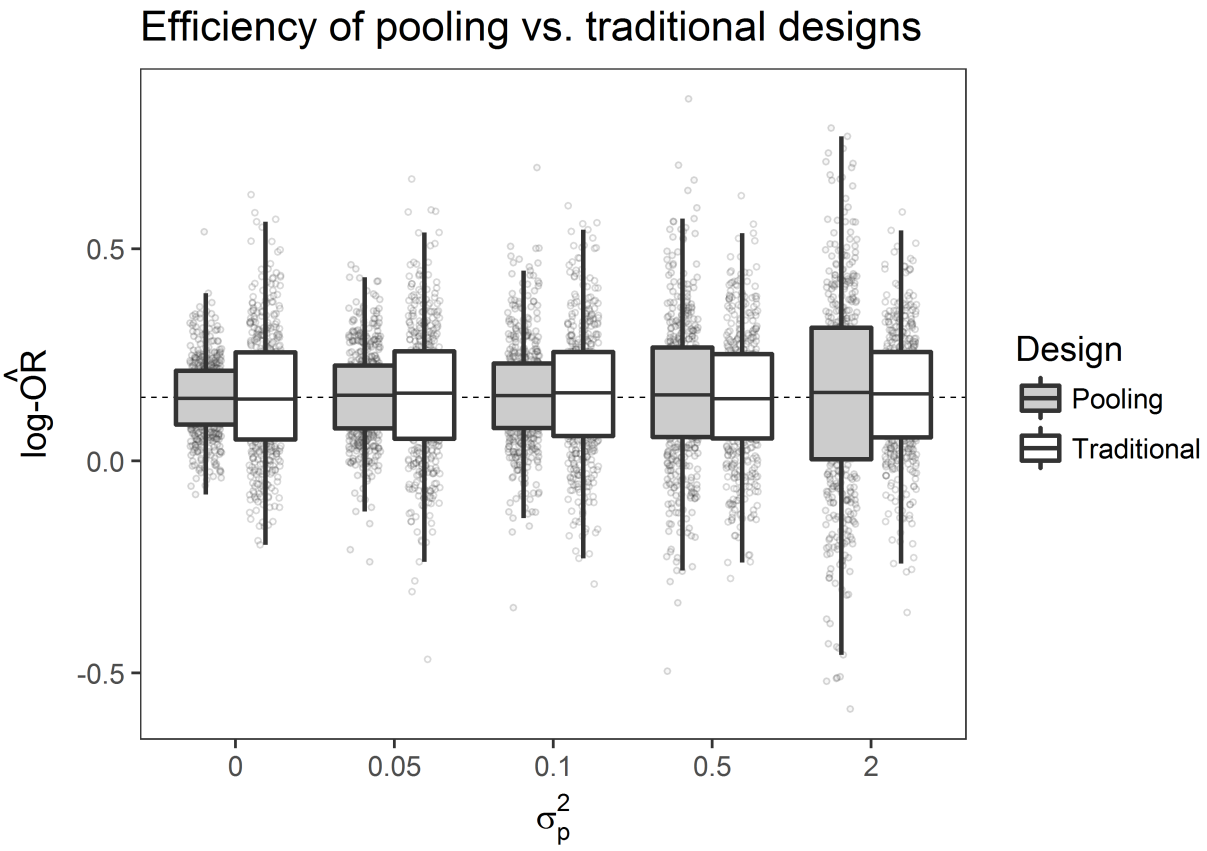

Figure 3.9: Boxplots of log-OR estimates for pooling and traditional designs (500 trials each, true log-OR = 0.15).

(1999) poolwise logistic regression model, or use one of the discriminant function approaches. However, while assay measurement error may be negligible in certain scenarios, we believe negligible processing error is a strong and seldom justifiable assumption. In our motivating example, the estimated processing error variance was indeed too large to ignore. Thus, we feel that performing valid inference with poolwise data will almost always require error modeling. Our Gamma-based methods extend prior approaches to accommodate skewed biomarkers, which tend to be much more common than normally distributed biomarkers.

Our methods are also not limited to the pooling scenario for which they were developed. A special case for all four methods is all $g_i = 1$, a traditional design with no pooling. Thus, our R functions apply to a wide range of scenarios for estimating exposure-disease

associations. They can handle pooling or traditional designs with or without covariates, for a normal or skewed exposure measured precisely or with errors (additive or multiplicative), incorporating replicates if available, and either assuming a constant odds ratio or allowing it to vary with exposure level and covariates.

One potential problem with the logistic regression methods is that they are based on a likelihood function that assumes prospective sampling. The $Y|(X, \boldsymbol{C})$ part is not problematic, given the Prentice and Pyke (1979) results, but the $X|\boldsymbol{C}$ model could be affected by case-oversampling. That is, even if the individual-level $X|\boldsymbol{C}$ model (linear regression for NLR, constant-scale Gamma for GLR) are correctly specified for the population, that relationship may not hold within cases and within controls, and thus may not hold in a case-control study where the proportion of cases is far higher than in the population. Guolo (2008) suggests that using the prospective likelihood is valid if the specified distribution for the error-prone covariate ($X|\boldsymbol{C}$ in our case) is correct in the case-control sampling scheme, which is intuitive. Specifying and assessing a model for an imperfectly measured exposure is typically one of the hardest parts of a measurement error correction (Carroll *et al.*, 2006). But a unique feature of the pooling context is that if there is processing error only, the singles are actually precisely measured, and thus the $X|\boldsymbol{C}$ model can be directly assessed with that data. So our logistic regression methods should be valid in case-control studies, provided the $X|\boldsymbol{C}$ model is supported by the data on hand, which can be directly assessed in certain cases. The discriminant function methods are based on models for $X|(Y, \boldsymbol{C})$ and are therefore unaffected by sampling rates for Y.

While both processing error and measurement error have the effect of reducing the efficiency advantage of a pooling design vs. a traditional design, processing error is particularly worrisome because it can render the pooling design counterproductive. In fact, this may have occurred in our motivating example. The GDFA model gave $\hat{\sigma}_p^2 = 0.62$, and in simulations mimicking the CPP data the pooling design was less efficient than traditional for $\sigma_p^2 \geq 0.5$

(Figure 3.9). Absent errors, pooling designs offer gains in statistical efficiency limited only by the number of samples that can be feasibly combined in the lab. With processing error, if σ_p^2 is large enough, the pooling design may be less efficient than traditional for the same number of assays regardless of how large the pools are. Adaptive study designs could be considered, whereby a pooling study is initiated, but a stopping rule is in place to transition to all $g = 1$ if it becomes clear that σ_p^2 is prohibitively large.

From a study design standpoint, in cases where obtaining replicates is feasible, we currently favor one large pool size in addition to singles with replicates, as opposed to three different pool sizes including singles (Schisterman *et al.*, 2010). The main reason is to avoid potential validity issues that could arise if the error variances are in fact not independent of pool size, which seems particularly plausible for processing errors. Additionally, when conditions are favorable for pooling (e.g. σ_p^2 not too large; per-assay costs >> per-subject costs), larger pools may be much more informative of parameters of interest. In a P-1-2-3 design, for example, replacing the pools of size 2 with pools of size 3 will often lead to improved power. The pools of size 2 serve a purpose—they help distinguish the variance terms—but replicates accomplish this more directly. There may be counterexamples where relying on different pool sizes is more cost-effective than including replicate singles and one large pool size, but we suspect these are rare.

In future applied work, it will be valuable to search for ways to minimize processing errors and determine whether certain types of biospecimens (blood, saliva, etc.) are more or less susceptible to these errors. On the statistical side, the assumption that the processing error variance is constant with pool size needs to be vetted and perhaps modified, as it seems likely that larger pools would have larger errors. This is a key assumption that affects identifiability requirements and efficiency results. Additionally, it would be useful to develop less parametric approaches for improved robustness, ideally relaxing distributional assumptions on the errors and not having to specify the exposure given covariates distribution.

Chapter 4: Future work

4.1 Propensity score calibration with multiple confounders

The PSC procedure introduced by Stürmer *et al.* (2005) is compelling, and it would seem to be extremely well-suited for handling multiple unmeasured confounders when validation data are internal. A likelihood approach would typically require assuming multivariate normality of the unmeasured confounders or specifying a series of models for each unmeasured confounder (Spiegelman *et al.*, 2000); multivariate versions of RC are available but require continuous variables (Rosner *et al.*, 1990). PSC would require its usual three models: a logistic regression for the error-prone propensity score, a logistic regression for the gold standard propensity score, and a linear model relating them. We would anticipate valid estimation via PSC if the surrogacy assumption is relaxed, which internal validation data permits. Further, bias and/or efficiency losses from the RC procedure could likely be avoided by performing PSC via ML, while sacrificing some computational convenience.

4.2 Conditional logistic regression with pooling

Pooling-related gains in statistical efficiency are very exciting, and we look forward to developing additional methods to ensure validity of pooling designs while perhaps even further increasing efficiency. One way to do this is to adapt our error correction methods to conditional logistic regression, where pools are still homogeneous with respect to case status, but each case pool is comprised of cases that are covariate-matched to members of a corresponding control pool. Saha-Chaudhuri *et al.* (2011) developed a conditional logistic regression model for pooling that could be modified to include measurement and processing errors. The main difference from our logistic regression methods is that the integral is two-dimensional rather than one-dimensional, because the unit of observation is each matched case/control pool, and the precise summed biomarker level is unobserved for both pools. Numerical integration will be time-consuming, but can be avoided in the additive normal errors case by using the probit approximation for the logistic-normal integral (Carroll *et al.*, 2006). Although not described fully here, we have implemented the approach in our R package **pooling**, with replicates also supported.

4.3 Paired t-test designs

Alternatively, we note that analyzing matched case-control data via a paired t-test rather than conditional logistic regression could motivate a simple and very powerful study design. This assumes that matching is done on all covariates, as opposed to having additional model covariates. With a single pool size, each pooled measurement would be subject to measurement and processing errors of the same magnitude. The usual paired t-test would be valid (Carroll *et al.*, 2006), with the additional errors simply detracting from the power of the test

relative to the errors being absent. There would seem to be numerous advantages to this approach: simplicity, by only requiring one pool size and permitting a standard t-test analysis; robustness, as the t-test would be valid regardless of the biomarker and error distributions (assuming a reasonable sample size); and efficiency, as it would not be necessary to use up assays on smaller pools or replicates to help distinguish variance terms. Notably, this design would permit estimating the difference in mean biomarker levels for cases and controls, but not the odds ratio.

4.4 Expanding suite of pooling functions

In future work, we plan to broaden the scope of our **pooling** package to include functions for analyzing a wider range of poolwise data. This is especially important for measurement error corrections, where it is not possible to utilize standard procedures like *PROC GLM* in SAS or *glm* in R with appropriately specified weights or offsets. We focused solely on logistic regression here, but similar methods could be implemented for continuous outcomes, e.g. for a linear regression or a constant-scale Gamma disease model. In either case, the variable subject to pooling and errors could be a predictor or the outcome. Pooling does not offer efficiency gains in all scenarios, but there are other reasons to utilize pooling, so it would be useful to accommodate as many scenarios as possible whether they correspond to efficient study designs or not.

4.5 Tools for designing pooling studies

We believe that developing tools for designing pooling studies, e.g. calculating sample size and choosing appropriate (maybe optimal) pool sizes, will help make the methods more

accessible. Such calculations are not straightforward, particularly for the hybrid design proposed by Schisterman *et al.* (2010) and utilized in our approaches. Even a seemingly conservative approach, like calculating n for 80% power for a traditional design and then deploying a pooling study with n total assays and an equal number of pools of size 1, 2, and 3, may not be justified. Depending on the variance of measurement errors (σ_m^2) and processing errors (σ_p^2), the pooling design may be less efficient than the traditional for those pool sizes, leading to inadequate power with n assays. This may or may not be a rare occurrence; it seemed to occur in our motivating example.

One approach is to again consider a matching design with a single pool size and a paired t-test analysis. In that scenario, it is easy to calculate sample size and power, but doing so requires specifying σ_m^2 and σ_p^2. The former may be feasible based on technical documentation from the assay manufacturer or prior literature, while σ_p^2 would be hard to predict. Perhaps a conservative choice could be something like one-half the variance of the biomarker level. But the specified values for σ_m^2 and σ_p^2 also dictate the optimal pool size, so a conservative choice for σ_p^2 may lead to a poor choice for the pool size. We are currently developing web apps for visualizing design aspects of pooling studies, which will hopefully help investigators make design choices while also anticipating potential consequences of larger than expected processing error.

Appendix: R code for motivating examples

Chapter 1

```
# Install and load meuc package
install_github("vandomed/meuc")
library("meuc")

# Fit corrective methods for Table 1.4 - note that main and ext are data
# frames containing EAGeR and BioCycle data, respectively

# 2-model ML (full)
ml2.full <- ml_logistic_linear(
  main = main,
  external = ext,
  y_var = "pregtest",
  z_var = "cal.100",
  d_vars = "height",
  c_vars = c("vitd.below30", "age", "overweight"),
  approx_integral = FALSE,
  control = list(trace = 1, rel.tol = 1e-9)
)

# 2-model ML (approximate)
ml2.approx <- ml_logistic_linear(
  main = main,
  external = ext,
  y_var = "pregtest",
  z_var = "cal.100",
```

```
  d_vars = "height",
  c_vars = c("vitd.below30", "age", "overweight"),
  approx_integral = TRUE,
  control = list(trace = 1, rel.tol = 1e-9)
)

# 3-model ML
fit.ml3 <- ml_logistic_logistic_linear(
  main = main,
  external = ext,
  y_var = "pregtest",
  x_var = "vitd.below30",
  z_var = "cal.100",
  d_vars = "height",
  c_vars = c("age", "overweight"),
  estimate_var = TRUE,
  control = list(trace = 1, rel.tol = 1e-9)
)

# Regression calibration
rc <- rc_algebraic(
  main = main,
  external = ext,
  y_var = "pregtest",
  z_var = "cal.100",
  d_var = "height",
  c_vars = c("vitd.below30", "age", "overweight"),
  tdm_family = "binomial"
)

# Propensity score calibration
psc <- psc_cond_exp(
  main = main,
  external = ext,
  y_var = "pregtest",
  x_var = "vitd.below30",
  gs_vars = c("cal.100", "age", "overweight"),
  ep_vars = c("age", "overweight"),
  tdm_family = "binomial",
  boot_var = TRUE, boots = 1000
)

# Propensity score calibration with D
```

```
psc.d <- psc_algebraic_d(
  main = main,
  external = ext,
  y_var = "pregtest",
  x_var = "vitd.below30",
  d_var = "height",
  gs_vars = c("cal.100", "age", "overweight"),
  ep_vars = c("age", "overweight"),
  tdm_family = "binomial",
  boot_var = TRUE, boots = 1000
)
```

Chapter 2

```
# Install and load pooling package
install_github("vandomed/pooling")
library("pooling")

# Fit models for Table 2.9 - note that cpp.df is a data frame containing
# CPP data, and mcp1_10x.reps is a list containing one MCP-1 measurement
# for pools without replicates and two MCP-1 measurements for pools with
# replicates

# LRF without replicates and neither error
lrf.woreps.neither <- p_logreg_xerrors(
  g = cpp.df$g,
  y = cpp.df$SA_onezero,
  xtilde = cpp.df$mcp1_10x,
  c = cpp.df[, c("m_age", "nw", "smoke")],
  errors = "neither",
  control = list(trace = 1, rel.tol = 1e-9)
)

# LRF without replicates and PE only
lrf.woreps.pe <- p_logreg_xerrors(
  g = cpp.df$g,
  y = cpp.df$SA_onezero,
  xtilde = cpp.df$mcp1_10x,
  c = cpp.df[, c("m_age", "nw", "smoke")],
  errors = "processing",
```

```
  approx_integral = FALSE,
  control = list(trace = 1, rel.tol = 1e-9)
)

# LRF without replicates and ME only
lrf.woreps.me <- p_logreg_xerrors(
  g = cpp.df$g,
  y = cpp.df$SA_onezero,
  xtilde = cpp.df$mcp1_10x,
  c = cpp.df[, c("m_age", "nw", "smoke")],
  errors = "measurement",
  approx_integral = FALSE,
  control = list(trace = 1, rel.tol = 1e-9)
)

# LRA without replicates and PE only
lra.woreps.pe <- p_logreg_xerrors(
  g = cpp.df$g,
  y = cpp.df$SA_onezero,
  xtilde = cpp.df$mcp1_10x,
  c = cpp.df[, c("m_age", "nw", "smoke")],
  errors = "processing",
  approx_integral = TRUE,
  control = list(trace = 1, rel.tol = 1e-9)
)

# LRA without replicates and ME only
lra.woreps.pe <- p_logreg_xerrors(
  g = cpp.df$g,
  y = cpp.df$SA_onezero,
  xtilde = cpp.df$mcp1_10x,
  c = cpp.df[, c("m_age", "nw", "smoke")],
  errors = "measurement",
  approx_integral = TRUE,
  control = list(trace = 1, rel.tol = 1e-9)
)

# DFA without replicates and neither error
dfa.woreps.neither <- p_dfa_xerrors(
  g = cpp.df$g,
  y = cpp.df$SA,
  xtilde = cpp.df$mcp1_10x,
  c = cpp.df[, c("m_age", "nw", "smoke")],
```

```
  errors = "neither",
  control = list(trace = 1, rel.tol = 1e-9)
)

# DFA without replicates and PE only
dfa.woreps.pe <- p_dfa_xerrors(
  g = cpp.df$g,
  y = cpp.df$SA,
  xtilde = cpp.df$mcp1_10x,
  c = cpp.df[, c("m_age", "nw", "smoke")],
  errors = "processing",
  control = list(trace = 1, rel.tol = 1e-9)
)

# DFA without replicates and ME only
dfa.woreps.me <- p_dfa_xerrors(
  g = cpp.df$g,
  y = cpp.df$SA,
  xtilde = cpp.df$mcp1_10x,
  c = cpp.df[, c("m_age", "nw", "smoke")],
  errors = "measurement",
  control = list(trace = 1, rel.tol = 1e-9)
)

# LRF with replicates and ME only
lrf.wreps.me <- p_logreg_xerrors(
  g = cpp.df$g,
  y = cpp.df$SA_onezero,
  xtilde = mcp1_10x.reps,
  c = cpp.df[, c("m_age", "nw", "smoke")],
  errors = "measurement",
  approx_integral = FALSE,
  control = list(trace = 1, rel.tol = 1e-9)
)

# LRF with replicates and both errors
lrf.wreps.both <- p_logreg_xerrors(
  g = cpp.df$g,
  y = cpp.df$SA_onezero,
  xtilde = mcp1_10x.reps,
  c = cpp.df[, c("m_age", "nw", "smoke")],
  errors = "both",
  approx_integral = FALSE,
```

```
  control = list(trace = 1, rel.tol = 1e-9)
)

# LRA with replicates and ME only
lra.wreps.me <- p_logreg_xerrors(
  g = cpp.df$g,
  y = cpp.df$SA_onezero,
  xtilde = mcp1_10x.reps,
  c = cpp.df[, c("m_age", "nw", "smoke")],
  errors = "measurement",
  approx_integral = TRUE,
  control = list(trace = 1, rel.tol = 1e-9)
)

# LRA with replicates and both errors
lra.wreps.me <- p_logreg_xerrors(
  g = cpp.df$g,
  y = cpp.df$SA_onezero,
  xtilde = mcp1_10x.reps,
  c = cpp.df[, c("m_age", "nw", "smoke")],
  errors = "both",
  approx_integral = TRUE,
  control = list(trace = 1, rel.tol = 1e-9)
)

# DFA with replicates and ME only
dfa.wreps.me <- p_dfa_xerrors(
  g = cpp.df$g,
  y = cpp.df$SA,
  xtilde = mcp1_10x.reps,
  c = cpp.df[, c("m_age", "nw", "smoke")],
  errors = "measurement",
  control = list(trace = 1, rel.tol = 1e-9)
)

# DFA with replicates and both errors
dfa.wreps.both <- p_dfa_xerrors(
  g = cpp.df$g,
  y = cpp.df$SA,
  xtilde = mcp1_10x.reps,
  c = cpp.df[, c("m_age", "nw", "smoke")],
  errors = "both",
  control = list(trace = 1, rel.tol = 1e-9)
```

```
)
```

Chapter 3

```
# Fit models for Table 3.13

# Naive logistic regression
naive <- p_logreg(
  g = cpp.df$g,
  y = cpp.df$SA_onezero,
  x = cpp.df[, c("mcp1_10x", "m_age", "nw", "smoke")]
)

# Normal logistic regression
nlr <- p_logreg_xerrors(
  g = cpp.df$g,
  y = cpp.df$SA_onezero,
  xtilde = mcp1_10x.reps,
  c = cpp.df[, c("m_age", "nw", "smoke")],
  errors = "both",
  approx_integral = FALSE,
  control = list(trace = 1, rel.tol = 1e-9)
)

# Gamma logistic regression
glr <- p_logreg_xerrors2(
  g = cpp.df$g,
  y = cpp.df$SA_onezero,
  xtilde = mcp1_10x.reps,
  c = c.list,
  errors = "both",
  control = list(trace = 1, rel.tol = 1e-9)
)

# Gamma logistic regression without replicates
glr.woreps <- p_logreg_xerrors2(
  g = cpp.df$g,
  y = cpp.df$SA_onezero,
  xtilde = cpp.df$mcp1_10x,
  c = c.list,
```

```
  errors = "both",
  control = list(trace = 1, rel.tol = 1e-9)
)

# Fit models for Table 3.14

# Normal discriminant function
ndfa <- p_dfa_xerrors(
  g = cpp.df$g,
  y = cpp.df$SA,
  xtilde = mcp1_10x.reps,
  c = cpp.df[, c("m_age", "nw", "smoke")],
  constant_or = TRUE,
  errors = "both",
  control = list(trace = 1, rel.tol = 1e-9)
)

# Gamma discriminant function
gdfa <- p_dfa_xerrors2(
  g = cpp.df$g,
  y = cpp.df$SA_onezero,
  xtilde = mcp1_10x.reps,
  c = c.list,
  constant_or = TRUE,
  errors = "both",
  control = list(trace = 1, rel.tol = 1e-9)
)

# Gamma discriminant function without replicates
gdfa.woreps <- p_dfa_xerrors2(
  g = cpp.df$g,
  y = cpp.df$SA_onezero,
  xtilde = cpp.df$mcp1_10x,
  c = c.list,
  constant_or = TRUE,
  errors = "both",
  control = list(trace = 1, rel.tol = 1e-9)
)

# Normal discriminant function with non-constant OR
ndfa.nonconstant <- p_dfa_xerrors(
  g = cpp.df$g,
  y = cpp.df$SA,
```

```
  xtilde = mcp1_10x.reps,
  c = cpp.df[, c("m_age", "nw", "smoke")],
  constant_or = FALSE,
  errors = "both",
  control = list(trace = 1, rel.tol = 1e-9)
)

# Gamma discriminant function with non-constant OR
gdfa.nonconstant <- p_dfa_xerrors2(
  g = cpp.df$g,
  y = cpp.df$SA_onezero,
  xtilde = mcp1_10x.reps,
  c = c.list,
  constant_or = FALSE,
  errors = "both",
  control = list(trace = 1, rel.tol = 1e-9)
)

# Log-OR vs. MCP-1 plots for Figure 3.8
library("ggplot2")

# Normal discriminant function
p.ndfa <- plot_dfa(
  estimates = ndfa.nonconstant$estimates,
  varcov = ndfa.nonconstant$theta.var,
  xrange = c(0.02, 6.75),
  xname = "MCP-1 (ng/mL) x10",
  cvals = list(c(26, 0, 0), c(26, 0, 1), c(26, 1, 0), c(26, 1, 1)),
  set_labels = c("NW = 0, Smoking = 0",
                 "NW = 0, Smoking = 1",
                 "NW = 1, Smoking = 0",
                 "NW = 1, Smoking = 1")) +
  labs(title = "Estimated log-OR from normal discriminant function") +
  theme_bw(base_size = 15)
plot(p.ndfa)

# Gamma discriminant function
p.gdfa <- plot_dfa2(
  estimates = gdfa.nonconstant$estimates,
  varcov = gdfa.nonconstant$theta.var,
  xrange = c(0.02, 6.75),
  xname = "MCP-1 (ng/mL) x10",
  cvals = list(c(26, 0, 0), c(26, 0, 1), c(26, 1, 0), c(26, 1, 1)),
```

```
  set_labels = c("NW = 0, Smoking = 0",
                 "NW = 0, Smoking = 1",
                 "NW = 1, Smoking = 0",
                 "NW = 1, Smoking = 1")) +
  labs(title = "Estimated log-OR from Gamma discriminant function") +
  theme_bw(base_size = 15)
plot(p.gdfa)
```

Bibliography

Abrevaya J, Hausman JA (2004). "Response error in a transformation model with an application to earnings-equation estimation." *The Econometrics Journal*, **7**(2), 366–388.

Akaike H (1974). "A new look at the statistical model identification." *IEEE transactions on automatic control*, **19**(6), 716–723.

Berntsen J, Espelid TO, Genz A (1991). "An adaptive algorithm for the approximate calculation of multiple integrals." *ACM Transactions on Mathematical Software (TOMS)*, **17**(4), 437–451.

Borchers HW (2017). *pracma: Practical Numerical Math Functions*. R package version 2.1.1, URL `https://CRAN.R-project.org/package=pracma`.

Camilli G (1995). "Origin of the scaling constant d = 1.7 in item response theory: correction."

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM (2006). *Measurement error in nonlinear models: a modern perspective*. CRC press.

Chen P, Tebbs JM, Bilder CR (2009). "Group testing regression models with fixed and random effects." *Biometrics*, **65**(4), 1270–1278.

Clayton D, *et al.* (1992). "Models for the analysis of cohort and case-control studies with inaccurately measured exposures." *Statistical models for longitudinal studies of health*, pp. 301–331.

Cornfield J, *et al.* (1962). "Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis." In *Fed Proc*, volume 21, pp. 58–61.

Dorfman R (1943). "The detection of defective members of large populations." *The Annals of Mathematical Statistics*, **14**(4), 436–440.

Efron B, Tibshirani R (1986). "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy." *Statistical science*, pp. 54–75.

Firth D (1993). "Bias reduction of maximum likelihood estimates." *Biometrika*, **80**(1), 27–38.

Frerichs RR, Srinivasan SR, Webber LS, Berenson G (1976). "Serum cholesterol and triglyceride levels in 3,446 children from a biracial community: the Bogalusa Heart Study." *Circulation*, **54**(2), 302–309.

Fuller W (1987). "Measurement error Models. 1987 John Wiley & Sons." *New York*, pp. 1–439.

Gaskins AJ, Mumford SL, Zhang C, Wactawski-Wende J, Hovey KM, Whitcomb BW, Howards PP, Perkins NJ, Yeung E, Schisterman EF (2009). "Effect of daily fiber intake on reproductive function: the BioCycle Study–." *The American journal of clinical nutrition*, **90**(4), 1061–1069.

Genz AC, Malik AA (1980). "Remarks on algorithm 006: An adaptive algorithm for numerical integration over an N-dimensional rectangular region." *Journal of Computational and Applied mathematics*, **6**(4), 295–302.

Greenland S (1988). "Variance estimation for epidemiologic effect estimates under misclassification." *Statistics in Medicine*, pp. 745–757.

Greenland S (2000). "An introduction to instrumental variables for epidemiologists." *International journal of epidemiology*, **29**(4), 722–729.

Greenland S, Pearl J, Robins JM (1999). "Causal diagrams for epidemiologic research." *Epidemiology*, pp. 37–48.

Guolo A (2008). "A flexible approach to measurement error correction in case–control studies." *Biometrics*, **64**(4), 1207–1214.

Hardy JB (2003). "The collaborative perinatal project: lessons and legacy." *Annals of epidemiology*, **13**(5), 303–311.

Kipnis V, Midthune D, Freedman LS, Carroll RJ (2012). "Regression calibration with more surrogates than mismeasured variables." *Statistics in medicine*, **31**(23), 2713–2732.

Kuha J (1994). "Corrections for exposure measurement error in logistic regression models with an application to nutritional data." *Statistics in medicine*, **13**(11), 1135–1148.

Lehmann EL, Casella G (2006). *Theory of point estimation*. Springer Science & Business Media.

Liu Y, McMahan C, Gallagher C (2017). "A general framework for the regression analysis of pooled biomarker assessments." *Statistics in medicine*, **36**(15), 2363–2377.

Lunt M, Glynn RJ, Rothman KJ, Avorn J, Stürmer T (2012). "Propensity score calibration in the absence of surrogacy." *American journal of epidemiology*, **175**(12), 1294–1302.

Lyles RH, Guo Y, Hill AN (2009). "A fresh look at the discriminant function approach for estimating crude or adjusted odds ratios." *The American Statistician*, **63**(4), 320–327.

Lyles RH, Kupper LL (1997). "A detailed evaluation of adjustment methods for multiplicative measurement error in linear regression with applications in occupational epidemiology." *Biometrics*, pp. 1008–1025.

Lyles RH, Kupper LL (2013). "Approximate and Pseudo-Likelihood Analysis for Logistic Regression Using External Validation Data to Model Log Exposure." *Journal of agricultural, biological, and environmental statistics*, **18**(1), 22–38.

Lyles RH, Mitchell EM, Weinberg CR, Umbach DM, Schisterman EF (2016). "An efficient design strategy for logistic regression using outcome-and covariate-dependent pooling of biospecimens prior to assay." *Biometrics*, **72**(3), 965–975.

Lyles RH, Van Domelen D, Mitchell EM, Schisterman EF (2015). "A discriminant function approach to adjust for processing and measurement error when a biomarker is assayed in pooled samples." *International journal of environmental research and public health*, **12**(11), 14723–14740.

Mendall M, Patel P, Ballam L, Strachan D, Northfield T (1996). "C reactive protein and its relation to cardiovascular risk factors: a population based cross sectional study." *Bmj*, **312**(7038), 1061–1065.

Mitchell EM, Lyles RH, Manatunga AK, Danaher M, Perkins NJ, Schisterman EF (2014a). "Regression for skewed biomarker outcomes subject to pooling." *Biometrics*, **70**(1), 202–211.

Mitchell EM, Lyles RH, Manatunga AK, Perkins NJ, Schisterman EF (2014b). "A highly efficient design strategy for regression with outcome pooling." *Statistics in medicine*, **33**(28), 5028–5040.

Mitchell EM, Lyles RH, Schisterman EF (2015). "Positing, fitting, and selecting regression models for pooled biomarker data." *Statistics in medicine*, **34**(17), 2544–2558.

Mumford S, Sjaarda L, Silver R, Radin R, Mitchell E, Schisterman E (2015). "C-reactive protein and pregnancy loss: results from the effects of aspirin in gestation and reproduction (EAGeR) trial." *Fertility and Sterility*, **3**(104), e352.

Narasimhan B, Johnson SG (2017). *cubature: Adaptive Multivariate Integration over Hypercubes*. R package version 1.3-11, URL `https://CRAN.R-project.org/package=cubature`.

Nemes S, Jonasson JM, Genell A, Steineck G (2009). "Bias in odds ratios by logistic regression modelling and sample size." *BMC medical research methodology*, **9**(1), 56.

Prentice RL, Pyke R (1979). "Logistic disease incidence models and case-control studies." *Biometrika*, **66**(3), 403–411.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

Rosner B, Spiegelman D, Willett W (1990). "Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error." *American journal of epidemiology*, **132**(4), 734–745.

Rosner B, Willett W, Spiegelman D (1989). "Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error." *Statistics in medicine*, **8**(9), 1051–1069.

Saha-Chaudhuri P, Umbach DM, Weinberg CR (2011). "Pooled exposure assessment for matched case-control studies." *Epidemiology (Cambridge, Mass.)*, **22**(5), 704.

Schisterman EF, Silver RM, Lesher LL, Faraggi D, Wactawski-Wende J, Townsend JM, Lynch AM, Perkins NJ, Mumford SL, Galai N (2014). "Preconception low-dose aspirin and pregnancy outcomes: results from the EAGeR randomised trial." *The Lancet*, **384**(9937), 29–36.

Schisterman EF, Silver RM, Perkins NJ, Mumford SL, Whitcomb BW, Stanford JB, Lesher LL, Faraggi D, Wactawski-Wende J, Browne RW, *et al.* (2013). "A Randomised Trial to Evaluate the Effects of Low-dose Aspirin in Gestation and Reproduction: Design and Baseline Characteristics." *Paediatric and perinatal epidemiology*, **27**(6), 598–609.

Schisterman EF, Vexler A, Mumford SL, Perkins NJ (2010). "Hybrid pooled–unpooled design for cost-efficient measurement of biomarkers." *Statistics in medicine*, **29**(5), 597–613.

Seber GA, Lee AJ (2012). *Linear regression analysis*, volume 329. John Wiley & Sons.

Smith AK, Ayanian JZ, Covinsky KE, Landon BE, McCarthy EP, Wee CC, Steinman MA (2011). "Conducting high-value secondary dataset analysis: an introductory guide and resources." *Journal of general internal medicine*, **26**(8), 920–929.

Spiegelman D, Carroll RJ, Kipnis V (2001). "Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument." *Statistics in medicine*, **20**(1), 139–160.

Spiegelman D, Rosner B, Logan R (2000). "Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs." *Journal of the American Statistical Association*, **95**(449), 51–61.

Stramer SL, Notari EP, Krysztof DE, Dodd RY (2013). "Hepatitis B virus testing by minipool nucleic acid testing: does it improve blood safety?" *Transfusion*, **53**(10pt2), 2449–2458.

Streeter AJ, Lin NX, Crathorne L, Haasova M, Hyde C, Melzer D, Henley WE (2017). "Adjusting for unmeasured confounding in nonrandomized longitudinal studies: a methodological review." *Journal of clinical epidemiology*, **87**, 23–34.

Stürmer T, Glynn RJ, Rothman KJ, Avorn J, Schneeweiss S (2007a). "Adjustments for unmeasured confounders in pharmacoepidemiologic database studies using external information." *Medical care*, **45**(10 SUPL), S158.

Stürmer T, Schneeweiss S, Avorn J, Glynn RJ (2005). "Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration." *American journal of epidemiology*, **162**(3), 279–289.

Stürmer T, Schneeweiss S, Rothman KJ, Avorn J, Glynn RJ (2007b). "Performance of propensity score calibration-a simulation study." *American journal of epidemiology*, **165**(10), 1110–1118.

Thurston SW, Spiegelman D, Ruppert D (2003). "Equivalence of regression calibration methods in main study/external validation study designs." *Journal of Statistical Planning and Inference*, **113**(2), 527–539.

Van Domelen DR (2018a). *meuc: Measurement Error and Unmeasured Confounding*. R package version 1.1.1, URL `https://github.com/vandomed/meuc`.

Van Domelen DR (2018b). *pooling: Fit Poolwise Regression Models*. R package version 1.1.1, URL `https://github.com/vandomed/pooling`.

VanderWeele TJ, Hernán MA, Robins JM (2008). "Causal directed acyclic graphs and the direction of unmeasured confounding bias." *Epidemiology (Cambridge, Mass.)*, **19**(5), 720.

Weinberg CR, Umbach DM (1999). "Using Pooled Exposure Assessment to Improve Efficiency in Case-Control Studies." *Biometrics*, **55**(3), 718–726.

Weinberg CR, Umbach DM (2014). "Correction to "Using pooled exposure assessment to improve efficiency in case-control studies," by Clarice R. Weinberg and David M. Umbach; 55, 718-726, September 1999." *Biometrics*, **70**(3), 1061.

Weller EA, Milton DK, Eisen EA, Spiegelman D (2007). "Regression calibration for logistic regression with multiple surrogates for one exposure." *Journal of Statistical Planning and Inference*, **137**(2), 449–461.

Whitcomb BW, Perkins NJ, Zhang Z, Ye A, Lyles RH (2012). "Assessment of skewed exposure in case-control studies with pooling." *Statistics in medicine*, **31**(22), 2461–2472.

Whitcomb BW, Schisterman EF, Klebanoff MA, Baumgarten M, Rhoton-Vlasak A, Luo X, Chegini N (2007). "Circulating chemokine levels and miscarriage." *American journal of epidemiology*, **166**(3), 323–331.

White E (2003). "Design and interpretation of studies of differential exposure measurement error." *American journal of epidemiology*, **157**(5), 380–387.

Wickham H, Hester J, Chang W (2018). *devtools: Tools to Make Developing R Packages Easier*. R package version 1.13.5, URL `https://CRAN.R-project.org/package=devtools`.

Zarek S, Plowden T, Silver R, Sjaarda L, Stanford J, Perkins N, DeCherney A, Mumford S, Schisterman E (2015). "Higher leptin levels are associated with decreased live birth: results from the effects of aspirin in gestation and reproduction (EAGeR) trial." *Fertility and Sterility*, **104**(3), e74.

Zhang X, Faries DE, Li H, Stamey JD, Imbens GW (2018). "Addressing unmeasured confounding in comparative observational research." *Pharmacoepidemiology and drug safety*, **27**(4), 373–382.