CLINICAL TRIALS *ARTICLE* *Clinical Trials* 2012; **9**: 721–729

Using audit information to adjust parameter estimates for data errors in clinical trials

Bryan E Shepherd[a], Pamela A Shaw[b] and Lori E Dodd[b]

Background Audits are often performed to assess the quality of clinical trial data, but beyond detecting fraud or sloppiness, the audit data are generally ignored. In an earlier study, using data from a nonrandomized study, Shepherd and Yu developed statistical methods to incorporate audit results into study estimates and demonstrated that audit data could be used to eliminate bias.

Purpose In this article, we examine the usefulness of audit-based error-correction methods in clinical trial settings where a continuous outcome is of primary interest.

Methods We demonstrate the bias of multiple linear regression estimates in general settings with an outcome that may have errors and a set of covariates for which some may have errors and others, including treatment assignment, are recorded correctly for all subjects. We study this bias under different assumptions, including independence between treatment assignment, covariates, and data errors (conceivable in a double-blinded randomized trial) and independence between treatment assignment and covariates but not data errors (possible in an unblinded randomized trial). We review moment-based estimators to incorporate the audit data and propose new multiple imputation estimators. The performance of estimators is studied in simulations.

Results When treatment is randomized and unrelated to data errors, estimates of the treatment effect using the original error-prone data (i.e., ignoring the audit results) are unbiased. In this setting, both moment and multiple imputation estimators incorporating audit data are more variable than standard analyses using the original data. In contrast, in settings where treatment is randomized but correlated with data errors and in settings where treatment is not randomized, standard treatment-effect estimates will be biased. And in all settings, parameter estimates for the original, error-prone covariates will be biased. The treatment and covariate effect estimates can be corrected by incorporating audit data using either the multiple imputation or moment-based approaches. Bias, precision, and coverage of confidence intervals improve as the audit size increases.

Limitations The extent of bias and the performance of methods depend on the extent and nature of the error as well as the size of the audit. This study only considers methods for the linear model. Settings much different than those considered here need further study.

Conclusions In randomized trials with continuous outcomes and treatment assignment independent of data errors, standard analyses of treatment effects will be unbiased and are recommended. However, if treatment assignment is correlated with data errors or other covariates, naive analyses may be biased. In these settings, and when covariate effects are of interest, approaches for incorporating audit results should be considered. *Clinical Trials* 2012; **9**: 721–729. http://ctj.sagepub.com

[a]Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, TN, USA, [b]Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, USA

Author for correspondence: Bryan E Shepherd, Department of Biostatistics, Vanderbilt University, 1161 21st Avenue South, S2323 MCN, Nashville, TN 37232-2158, USA.
Email: bryan.shepherd@vanderbilt.edu

10.1177/1740774512450100

Introduction

Clinical trials are subjected to quality control measures to ensure the validity of data and, hence, the accuracy of the study findings. Data audits are routinely performed, in which auditors compare the data sent to the data coordinating center with that in the source documents. Site-specific audits may be performed at a given frequency (e.g., every 3 years) to verify site quality, or trial-specific auditing procedures may be routinely implemented. The selection of a relatively small random sample provides assurance of data quality. If error rates are unacceptable, site enrollment may be suspended until the data-quality procedures are in place. However, beyond that, information from the audit about errors is typically ignored.

In a recent article, Shepherd and Yu [1] developed methods to incorporate audit findings into study estimates. Their methods are similar to those developed to adjust parameter estimates for classical covariate measurement error [2,3] but addressed a more general error structure where both the predictor and outcome variables could have errors, possibly correlated. Shepherd and Yu focused on the setting of a nonrandomized experiment and applied their methods to data from a multicenter observational cohort study of patients infected with HIV. Unlike the clinical trial setting, data audits are rare in observational studies, yet rates and magnitudes of errors are often quite high [4]. Shepherd and Yu found that, in certain settings, audit data could be used to reduce bias and to improve the precision of results. In this article, we will discuss incorporating audit findings to the analysis of data from clinical trials.

In randomized clinical trials (RCTs), the treatment effect is generally the parameter of primary interest. If there are important prognostic baseline covariates, it may be preferable to estimate the treatment effect from a model that adjusts for these covariates rather than performs a simple comparison of outcome between treatment groups. In some trials, randomization is stratified by important covariates; in these trials, analysts may also adjust for the stratifying covariates. Adjusting for covariates that are correlated with the outcome can improve the precision of the treatment-effect estimates and reduce bias due to chance imbalance in treatment assignment [5,6]. There is considerable debate over when and for which covariates to adjust for in a randomized trial, one concern being the bias that can be introduced when such covariates are chosen using a post hoc selection procedure [6–8]. We put this debate aside and presume a setting where important prognostic pretreatment covariates are known a priori to exist, are possibly recorded with errors, and an analysis adjusting for these covariates is of interest, either as a primary or supportive analysis of the treatment effect.

In the context of clinical trials, two types of covariate errors could be revealed by the auditing process: errors involving the treatment assignment and errors involving other participant data. We focus on the latter. Here, we consider the potential impact that errors in the recorded outcome or important baseline covariates have on the adjusted estimate of the treatment effect. We then consider the usefulness of audit data in this setting for improving estimation in the presence of these errors.

Data errors and implications for clinical trials

Model framework

Let Z be a $p \times 1$ vector of accurately measured covariates, including the treatment assignment, X is a $q \times 1$ vector of other important baseline covariates that may be observed with error, and Y is a continuous outcome of interest. Assume (Y,X,Z) follow the linear model

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$$

In the special case when Z is a univariate variable designating treatment assignment, this model is frequently referred to as an analysis of covariance (ANCOVA).

Instead of containing X and Y for all subjects, the trial database may contain observations for some subjects that are incorrect. To indicate this, we introduce W and Y^*, where

$$W = X + SU$$
$$Y^* = Y + S^y U^y + SU^*$$

with S being an indicator of the presence of an error in W, U representing the magnitude of the errors in W, U^* representing errors in Y^* induced by errors in W, S^y indicating other errors in Y^* unrelated to the errors in W, and U^y representing the magnitude of these other errors in Y^*. This model was introduced by Shepherd and Yu in the context of a cross-sectional HIV study where errors in the database for X (date of starting therapy in their example) may have induced errors in the recorded Y (CD4 count at start of therapy) [1]. This model allows for dependence in the errors of W and Y^* through (S,U,U^*), but assumes that (S,U,U^*) are independent of (S^y,U^y). Furthermore, we assume that (S,U) are independent of X, that (S^y,U^y,S,U^*) are independent of Y, and that ε is independent of (X,Z,S,U,S^y,U^y,U^*). These assumptions presume that database errors are clerical mistakes,

or something of that nature, that are not influenced by patient attributes or outcomes.

In a clinical trial, the data collection errors may be site dependent in a multisite trial or even related to the treatment assignment, say in an unblinded trial. Thus, we present methods, which allow for dependence between Z and the errors in W and Y^*. We will also give further consideration to special cases of error and covariate dependencies that are likely to arise in RCTs. Notice that if there are no errors in Y^* (i.e., $S^y = 0$ and $U^* = 0$ with probability one; hence $Y^* = Y$), then this model reduces to the classical measurement error problem [2,3], except that the distribution of errors in the recorded X is a mixture distribution with a point mass at zero.

For a more general setting than a randomized trial, Shepherd and Yu considered the impact of the data errors on estimation $\beta = (\beta_1,\beta_2)$ when one naively fits the linear model

$$E\left(Y^* \middle| W,Z\right) = \gamma_0 + \gamma_1 W + \gamma_2 Z$$

The parameters (γ_1,γ_2) in the naive model can be written as

$$\begin{pmatrix}\gamma_1\\ \gamma_2\end{pmatrix} = \begin{pmatrix}\Sigma_{ww} & \Sigma_{wz}\\ \Sigma_{zw} & \Sigma_{zz}\end{pmatrix}^{-1}\begin{pmatrix}\Sigma_{wy^*}\\ \Sigma_{zy^*}\end{pmatrix} \quad (1)$$

where Σ_{ab} represents the covariance of A and B. In contrast, the true parameters are given as

$$\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} = \begin{pmatrix}\Sigma_{xx} & \Sigma_{xz}\\ \Sigma_{zx} & \Sigma_{zz}\end{pmatrix}^{-1}\begin{pmatrix}\Sigma_{xy}\\ \Sigma_{zy}\end{pmatrix} \quad (2)$$

Let $T = W - X$ and $\tilde{T} = Y^* - Y$. It follows that $\Sigma_{ww} = \Sigma_{xx} + \Sigma_{tt}, \Sigma_{wz} = \Sigma_{xz} + \Sigma_{tz}, \Sigma_{wy^*} = \Sigma_{xy} + \Sigma_{x\tilde{t}} + \Sigma_{yt} + \Sigma_{t\tilde{t}}$ and $\Sigma_{zy^*} = \Sigma_{zy} + \Sigma_{z\tilde{t}}$. Therefore, estimates of (γ_1,γ_2) will generally be biased.

Review of methods

With an audit, the error, or lack thereof, in the outcome, $Y^* - Y$, and covariates, $W - X$, is observed for a subset of individuals in the study. This information can be used to correct for error-induced bias in the study estimates. Let V be the indicator that a subject is selected for an audit, and for $V = 1$, the true observations for X and Y are obtained. For notational simplicity, we assume that the audit is a simple random sample of subjects (Shepherd and Yu provide some discussion of how the methods can be extended in a straightforward manner to allow for dependence of V on Z, such as simple random sampling within sites).

Shepherd and Yu presented moment estimators that correct the naive ordinary least squares (OLS) estimators $\hat{\gamma} = (\hat{\gamma}_1,\hat{\gamma}_2)$ for the setting described in section 'Model framework', with the additional assumption that Z can only be correlated with the error in W(i.e., S,U) and Z cannot shift the mean of U. A more general version of their estimators, which is consistent without these added assumptions is

$$\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix} = \begin{pmatrix}\hat{\Sigma}_{ww} - \hat{\Sigma}_{tt} & \hat{\Sigma}_{wz} - \hat{\Sigma}_{zt}\\ \hat{\Sigma}_{zw} - \hat{\Sigma}_{zt} & \hat{\Sigma}_{zz}\end{pmatrix}^{-1}\begin{pmatrix}\hat{\Sigma}_{wy^*} - \hat{\Sigma}_{x\tilde{t}} - \hat{\Sigma}_{ty} - \hat{\Sigma}_{t\tilde{t}}\\ \hat{\Sigma}_{zy^*} - \hat{\Sigma}_{z\tilde{t}}\end{pmatrix} \quad (3)$$

One can obtain all the necessary moment estimates in equation (3) from the audit data, including $\hat{\Sigma}_{tt} = \widehat{Var}(W - X)$, $\hat{\Sigma}_{t\tilde{t}} = \widehat{Cov}(W - X, Y^* - Y)$, $\hat{\Sigma}_{z\tilde{t}} = \widehat{Cov}(Z, Y^* - Y)$, $\hat{\Sigma}_{x\tilde{t}} = \widehat{Cov}(X, Y^* - Y)$, and $\hat{\Sigma}_{ty} = \widehat{Cov}(Y, W - X)$ Alternate estimation approaches can be applied by setting some parameters as zero based on model assumptions or by estimating other parameters such as $P(S = 1)$ and $Var(U)$ [1]. Shepherd and Yu demonstrated how to construct confidence intervals (CIs) for β_1 in the univariate case using M-estimation techniques and large-sample theory; CIs for (β_1,β_2) in the multivariable case can be similarly constructed.

It should be recognized that these moment-based estimators assume that the original error-prone data were used as (W,Y^*), not a partially corrected version (with (W,Y^*) set as the truth for the audited records). If the data have been corrected (so that (W,Y^*) consists of (X,Y) for the audited records), then the moment-based estimators need to be slightly adjusted. When the audit is a simple random sample of all records, this can be done simply by multiplying all of the estimated variances for the error terms in equation (3) by one minus the proportion of records that were audited (i.e., multiplying $\hat{\Sigma}_{tt}$, $\hat{\Sigma}_{t\tilde{t}}$, $\hat{\Sigma}_{zt}$, and $\hat{\Sigma}_{x\tilde{t}}$ by $(N - n_v)/N$). This properly accounts for the fact that the variance of the data errors in the partially corrected data has been reduced. This procedure can be easily adapted to handle situations where the probability of being audited depends on Z.

An alternative approach for obtaining consistent estimators of (β_1,β_2) is to use techniques for dealing with missing data such as multiple imputation [9]. The complete data (Z,X,Y,W,Y^*,V) are only known for the subset of records that were audited, whereas only (Z,W,Y^*,V) is known for those records not audited. A multiple imputation approach that uses relationships observed in the complete data to impute is outlined in the Appendix. One benefit of using a missing data/multiple imputation approach is that it provides a natural way to account for missing data and two other types of errors that are often

observed in an audit: unverifiable values and unrecorded values. Specifically, an audit of the source documents may be unable to find some values that are recorded in the database, or the audit may discover that a value that was missing in the database was actually recorded in the source documents. In the former case, X or Y is missing, despite the fact that $V = 1$; in the latter case, (X,Y) are known, but either W or Y^* is missing.

Randomized trials

Implications for bias

For ease of notation, assume Z is the univariate treatment variable. In the special case of an RCT, randomization implies that Z will be independent of the baseline variables. Thus, it is reasonable to assume that Z is independent of X, S, U and, hence, W. Because Y^* is observed post baseline, there is the potential for the errors in observing Y to be correlated with Z. Nevertheless, in a double-blinded trial, it is reasonable to assume that all error terms in Y^* are also independent of Z. With these assumptions, one has $Cov(Z,W) = Cov(Z,X)$ and $Cov(Z,Y^*) = Cov(Z,Y)$. Using formulas (1) and (2), we can see that in this special case, no bias will occur in the usual least squares estimator for the treatment-effect parameter β_2 when one regresses Y^* on (W,Z). That is

$$E\left(Y^* \middle| W,Z\right) = \gamma_0 + \gamma_1 W + \beta_2 Z$$

In the simple case of a two-arm trial, this same result is perhaps more easily seen from the fact that the regression estimator of the treatment effect from the ANCOVA model (4) is

$$\hat{\beta}_2 = \bar{Y}_T^* - \bar{Y}_C^* + \hat{\gamma}_1\left(\bar{W}_T - \bar{W}_C\right)$$

where $\bar{W}_Z$ and $\bar{Y}_Z^*$ are the means of W and Y^* in treatment arms $Z = T$ (treatment) and $Z = C$ (control). Regardless of the frequency and magnitude of errors in W, randomization will tend to balance the covariate W between treatment groups, and thus, the error in W is not expected to introduce bias into the estimator of β_2. This result, and its implications for ANCOVA in the presence of covariate measurement error in linear models, is well known and is in sharp contrast to the impact of errors in W on $\hat{\beta}_2$ for a nonrandomized study [10].

It is important to note that this bias is a large-sample result and does not rule out chance imbalances between arms in the error-prone covariate W having an impact on the estimator of β_2 in double-blinded RCTs. Also, though errors in (Y^*,W) are not expected to introduce bias into the estimator of β_2 in a typical double-blinded randomized study, the error in these variables will have a negative impact on precision. It is interesting to note that in the case where there are errors only in W, say in trials where an end point review committee meticulously verifies every primary end point, this lack of precision generally cannot be ameliorated by estimating β_2 with methods that adjust for the error in W. In this case, the naive ANCOVA estimate for β_2 will provide the more efficient estimator of β_2 over a large class of measurement error methods [11]. Under more general error structures, Shepherd and Yu observed that error-corrected methods may not result in a better mean-squared error (MSE) unless the audit comprises a substantial proportion of study records. Thus, in the special case of randomized trials, fitting the simple ANCOVA model and not correcting for data errors in W will frequently yield the best estimator for β_2, in terms of MSE. In contrast, estimates of β_1 will be biased even in a double-randomized trial, so incorporating the audit data can improve estimation of β_1.

One could consider the case where, through incomplete blinding, Z could in some way be correlated with database errors made in the observation of Y^*. This could occur, for example, if a specific intervention led to a more or less error-prone process for end point documentation. In practice, this may occur with randomization to treatments from different specialties (e.g., surgery vs. nonsurgery), with different specialists tending to record data differently. Consider the case that $Cov(Z,S^yU^y) \neq 0$, but randomization yields independence of Z with (X,SU,SU^*). The standard OLS estimators for Y^* regressed on W,Z estimate

$$\gamma_2 = \Sigma_{zz}^{-1}\Sigma_{ww}^{-1}\left(\Sigma_{zy^*}\Sigma_{ww} - \Sigma_{wz}\Sigma_{wy^*}\right)$$

which implies $\gamma_2 - \beta_2 = \Sigma^{-1}{}_{zz}\,\Sigma_{zt^y}$, where $T^y = S^yU^y$. Thus, the treatment-effect parameter is biased, and it may be important to incorporate the audit data into estimates. Estimators of β_1 will also be biased.

Numeric example and simulation

In this section, we illustrate the bias induced by data errors under different assumptions and briefly investigate the performance of our estimators using audit data. Our simulation setting is based on that described by Shepherd and Yu, which was roughly set up to approximate the relationship between CD4 cell count (X) and $\log_{10}$-transformed HIV-1 RNA (Y) among HIV-infected adults. We extend this setting by including a treatment indicator (Z).

Specifically, let Z be from a Bernoulli distribution with success probability 0.5. Given Z, X is normally distributed with mean $200 + \mu_{xz}Z$ and variance 50^2. Given X and Z, Y is normally distributed with mean $\beta_0 + \beta_1 X + \beta_2 Z$ and variance 0.5^2, where $(\beta_0, \beta_1, \beta_2) = (6, -0.01, 1)$. W and Y^* follow the models given in section 'Model framework': S is from a Bernoulli distribution with success probability P; U is normally distributed with mean 0, variance σ_u^2; S^y is from a Bernoulli distribution with success probability P_y; U^y given Z is normally distributed with mean $\mu_{zu}{}^y Z$, variance 0.5^2; U^* is normally distributed with mean 0 and variance 0.5^2; and the correlation between U and U^* is given by ρ_{u,u^*}.

In Table 1, we vary the values of $\mu_{xz} \in \{-50,0\}$, $\mu_{zu}{}^y \in \{0,1\}$, $\rho_{u,u^*} \in \{0,0.5\}$, $\rho \in \{0.05,0.20\}$, $\sigma_u \in \{25,50\}$, and $P_y \in \{0.05,0.2\}$ to examine the bias of (γ_1,γ_2) with respect to (β_1,β_2) when the model $Y^* = \gamma_0 + \gamma_1 W + \gamma_2 Z$ is fit. By setting $\mu_{xz} = 0$, we mimic a randomized trial where Z is independent of X. If Z is also independent of the errors in Y^* (i.e., in our setup $\mu_{zu}{}^y = 0$), then γ_2 is unbiased for β_2. In contrast, γ_1 is a biased estimator of β_1, and the magnitude of the bias depends on the error rate in the recording of X (P) and the variance of the magnitude of the errors (σ_u) relative to the variance of X. In the randomized trial setting, if the magnitude of errors in Y^* depends on Z (i.e., in our setup $\mu_{zu}{}^y = 1$), then γ_2 is biased. Finally, in a nonrandomized trial setting where Z and X are dependent (i.e., $\mu_{xz} = -50$ in our setup), both γ_1 and γ_2 are biased, with the magnitude of the bias depending on the rate, magnitude, and correlation of errors.

We performed a limited set of simulations to examine the performance of estimators incorporating audit data. For these simulations, we generated data varying $(\mu_{xz},\mu_{zu}{}^y)$between (A) (0,0), a double-blinded randomized trial; (B) (0,1), a randomized trial with errors dependent on Z; and (C) (–50,1), an observational study with errors dependent on Z. The parameters $(\rho_{u,u^*},\sigma_u,P,P_y)$ were set as (0.5,50,0.2,0.2), resulting in the settings that corresponded to the maximum bias shown in Table 1. In each simulation, we generated $N = 1000$ vectors $(Y,X,Y^*,W,S,U,S^y,U^y,U^*,\varepsilon)$. Within each simulation experiment, we computed estimates with the number of randomly audited charts (n_v) being 0, 25, 50, 100, 300, 500, and 1000. When no charts were audited, we computed the naive estimates regressing Y^* on (W,Z). A total of 1000 audited charts corresponded to having correct data for all records and regressing Y on (X,Z). When the number of audited charts was between 25 and 500, we estimated (β_1,β_2) and computed 95% CIs using both the moment-based and multiple imputation estimators mentioned above; for those records not sampled for the audit (i.e., $V = 0$), we treated (Y,X,S,U,S^y,U^y,U^*) as if they were unknown. For each of the data-generating scenarios, we performed 1000 simulation replications. A second set of simulations were identical except $N = 100$, and n_v was either 0, 50, or 100. Simulation code is provided in the Supplementary Material (posted at http://biostat.mc.vanderbilt.edu/DataAuditSimulationCode2).

Table 2 shows simulation results for $N = 1000$. Under scenario A (double-blinded randomized trial), the naive estimate of β_2 had negligible bias and outperformed all moment-based, audit-corrected estimators in terms of MSE except for the estimator obtained after auditing all 1000 records and plugging in the true values for (Y,X,Z). With an audit of $n_v = 500$, there was little difference in performance between the multiple imputation and naive estimators. Under scenario B (unblinded randomized trial), as expected, the naive estimate of β_2 was biased, the MSE was large, and coverage was very poor (0.005). With as few as 25 audits, both the multiple imputation and moment estimators were essentially unbiased, but they were quite variable – particularly the moment-based estimator. An audit of $n_v = 50$ was needed for the moment-based estimator to outperform the naive estimator in terms of MSE, and only with an audit of $n_v = 500$ did the moment-based 95% CI cover at their nominal level. In contrast, with an audit of as few as 25 records, the multiple imputation approach resulted in a lower MSE than the naive estimator. The coverage appeared nominal after 50 audits. Similar trends, although with more extreme bias, were seen under scenario C (observational study). In all three scenarios, naive estimates of β_1 were biased; as the audit size increased, MSE decreased, and coverage improved.

These simulation results clearly demonstrate that with the error rates and magnitudes considered here, an audit of $n_v = 25$ is insufficient to provide corrected estimates with reasonable MSE. This is particularly notable with the moment-based estimators under scenario C. This result, however, is somewhat to be expected as with small audits, only a handful of audited records will have errors and will be used to estimate variances and covariances between errors.

In general, and throughout all simulations, the multiple imputation estimators tended to outperform their moment-based counterparts. The multiple imputation estimators reported in Table 2 drew from a fitted distribution where we (correctly) assumed errors were normally distributed. However, this correct model specification does not appear to be driving the superior performance of the multiple imputation estimators. In an additional set of simulations, we made no such normality assumption, instead imputing with the fitted value plus a random residual. The performance of this alternative

Table 1. Bias for β_1 and β_2 as a function of different error model parameters

Error parameters				Treatment randomized ($\mu_{xz} = 0$)		Treatment not randomized ($\mu_{xz} = -50$)	
ρ_{u,u^*}	σ_u	P	P_y	% Bias β_1	% Bias β_2	% Bias β_1	% Bias β_2
Errors in Y^* independent of Z (μ_{zu}^y) = 0							
0	25	0.05	0.05	–1.2	0	–1.2	0.6
			0.2	–1.2	0	–1.2	0.6
		0.2	0.05	–4.8	0	–4.8	2.4
			0.2	–4.8	0	–4.8	2.4
	50	0.05	0.05	–4.8	0	–4.8	2.4
			0.2	–4.8	0	–4.8	2.4
		0.2	0.05	–16.7	0	–16.7	8.4
			0.2	–16.7	0	–16.7	8.4
0.5	25	0.05	0.05	–2.5	0	–2.5	1.2
			0.2	–2.5	0	–2.5	1.2
		0.2	0.05	–9.5	0	–9.5	4.8
			0.2	–9.5	0	–9.5	4.8
	50	0.05	0.05	–7.1	0	–7.1	3.6
			0.2	–7.1	0	–7.1	3.6
		0.2	0.05	–25	0	–25	12.5
			0.2	–25	0	–25	12.5
Errors in Y^* dependent on Z (μ_{zu}^y) = 1							
0	25	0.05	0.05	–1.2	5	–1.2	5.6
			0.2	–1.2	20	–1.2	20.6
		0.2	0.05	–4.8	5	–4.8	7.4
			0.2	–4.8	20	–4.8	22.4
	50	0.05	0.05	–4.8	5	–4.8	7.4
			0.2	–4.8	20	–4.8	22.4
		0.2	0.05	–16.7	5	–16.7	13.3
			0.2	–16.7	20	–16.7	28.4
0.5	25	0.05	0.05	–2.5	5	–2.5	6.2
			0.2	–2.5	20	–2.5	21.2
		0.2	0.05	–9.5	5	–9.5	9.7
			0.2	–9.5	20	–9.5	24.8
	50	0.05	0.05	–7.1	5	–7.1	8.5
			0.2	–7.1	20	–7.1	23.6
		0.2	0.05	–25	5	–25	17.6
			0.2	–25	20	–25	32.5

ρ_{u,u^*} : correlation between errors in outcome and covariate; σ_u^2: variance of covariate errors; P: probability of covariate error; P_y: probability of error in outcome.

multiple imputation approach was similar to that of the parametric multiple imputation presented in Table 2 (see Table in Supplementary Material).

Table 3 demonstrates the performance of estimators under the same setup except with a smaller sample size ($N = 100$) and an audit of size $n_v = 50$. In Table 3, we have also included results if one were to analyze the data after replacing (Y^*,W) with (Y,X) for those n_v records that were audited, but leaving (Y^*,W) in those $N–n_v$ records that were not audited. This approach is common in practice, where the errors discovered by the audit are often fixed, but the unaudited data are treated as if they were correct. In Table 3, we label this as the naive correction. This naive correction outperforms the more naive analysis of regressing Y on W,Z; this is not surprising as one would expect bias to decrease by $(n_v/N) \times 100\%$ (which is seen in our simulations). However, except for the estimate of β_2 under scenario A (randomized trial with Z independent of X and errors), with $n_v = 50$, our moment-based and multiple imputation estimators outperformed the naive correction estimators.

Table 2. Performance of estimators under error scenarios relevant for different types of clinical studies

Scenario	N	n_v	Estimates of β_2						Estimates of β_1					
			Moment based			Multiple imputation			Moment based			Multiple imputation		
			% Bias	Coverage	MSE x 10^3	% Bias	Coverage	MSE x 10^3	% Bias	Coverage	MSE x 10^7	% Bias	Coverage	MSE x 10^7
A	1000	0[a]	0.31	0.951	1.7	0.31	0.951	1.7	−24.8	0	63.7	−24.8	0	63.7
		25	0.03	0.736	46	0.74	0.943	18.7	5.07	0.828	155	−2.54	0.899	28.7
		50	0.2	0.765	17.5	0.16	0.949	9.5	1.25	0.861	50.9	−1.16	0.894	14.3
		100	−0.21	0.798	8.7	0.01	0.928	5.3	0.97	0.914	23.6	−0.47	0.906	7.2
		300	0	0.886	3	0	0.944	2.1	0.34	0.951	6.9	−0.13	0.936	2.5
		500	0.19	0.931	1.8	0.12	0.952	1.5	0.07	0.954	3.7	−0.01	0.951	1.6
		1000[b]	0.18	0.946	1	0.18	0.946	1	0.01	0.96	1	0.01	0.96	1
B	1000	0[a]	19.8	0.005	41	19.8	0.005	41	−25	0	64.7	−25	0	64.7
		25	−0.39	0.634	65.9	0.22	0.906	24.8	5.61	0.862	223	−2.13	0.902	34.6
		50	−0.35	0.675	25.4	0.19	0.945	11.5	1.75	0.894	61.4	−1.1	0.919	14.7
		100	−0.08	0.746	11.4	0.03	0.939	6	1.23	0.933	27.1	−0.19	0.932	7.2
		300	−0.38	0.887	3.4	−0.27	0.948	2.2	0.83	0.936	8.9	0.13	0.922	3
		500	−0.23	0.942	2	−0.24	0.949	1.4	0.39	0.942	4.6	0.05	0.941	1.8
		1000[b]	−0.23	0.965	0.9	−0.23	0.965	0.9	−0.01	0.948	1	−0.01	0.948	1
C	1000	0[a]	32.7	0	109	32.7	0	109	−25.3	0	66	−25.3	0	66
		25	−19.3	0.754	15393	1.81	0.915	31.6	25.6	0.892	22,942	−1.81	0.903	33.2
		50	−2.2	0.801	74.5	1.15	0.927	15.2	3.92	0.931	107	−0.81	0.919	15.1
		100	−0.73	0.858	24	0.36	0.937	7.7	1.73	0.96	29.3	−0.19	0.933	7
		300	0.05	0.905	6.7	0.22	0.938	3.1	0.07	0.976	8	−0.12	0.95	2.5
		500	0.11	0.947	3.6	0.16	0.94	2	−0.15	0.967	4.3	−0.18	0.951	1.6
		1000[b]	0.13	0.948	1.3	0.13	0.948	1.3	−0.1	0.948	1	−0.1	0.948	1

A: double-blinded randomized trial. B: unblinded randomized trial. C: nonrandomized study. Details in text.
[a]No correction: linear model fit to original error-prone data.
[b]Complete correction: All 1000 records audited and linear model fit to corrected data.

Table 3. Performance of estimators under error scenarios relevant for different types of clinical studies in smaller samples ($N = 100$)

Scenario	Estimators[a]	Audit size	Estimates of β_2			Estimates of β_1		
			% Bias	Coverage	MSE x 10^3	% Bias	Coverage	MSE x 10^7
A	No correction	0	0	0.958	17.2	–24.7	0.471	81.8
	Naive correction	50	0.03	0.952	14.0	–13.2	0.753	34.1
	Moment based	50	0.12	0.935	20.4	2.8	0.948	49.3
	Multiple imputation	50	–0.19	0.944	15.1	0.0	0.931	18.2
	Complete correction	100	–0.25	0.934	10.6	0.8	0.939	10.9
B	No correction	0	19.9	0.732	59.7	–24.6	0.531	83.5
	Naive correction	50	9.6	0.899	24.4	–13.5	0.765	37.6
	Moment based	50	–0.56	0.927	23.0	2.1	0.947	51.8
	Multiple imputation	50	–0.10	0.95	15.8	–0.8	0.943	17.5
	Complete correction	100	–0.42	0.953	10.0	0.3	0.946	9.6
C	No correction	0	32.4	0.447	131	–23.8	0.552	80.6
	Naive correction	50	16.8	0.780	48.6	–12.7	0.788	35.9
	Moment based	50	–1.49	0.954	44.9	3.2	0.945	65.2
	Multiple imputation	50	0.4	0.958	19.4	–0.4	0.935	18.8
	Complete correction	100	0.18	0.956	12.5	0.2	0.944	10.3

A: double-blinded randomized trial. B: unblinded randomized trial. C: nonrandomized study. Details in text.
[a]No correction: Linear model fit to original error-prone data. Naive correction: data corrected for 50 records that were audited and linear model fit to complete data including both corrected and uncorrected. Moment based: estimator given in section 'Review of methods'. Multiple imputation: estimator given in Appendix. Complete correction: all 100 records audited and linear model fit to corrected data.

Discussion

Audits are commonly used to monitor a trial's data operations and are viewed primarily as a quality control and assurance measure. With the methods presented here, audits can also be seen as an analytical tool that allows for the correction of bias in trial estimates induced by database errors. We presented error-correction methods for general cases, where complex relationships could exist between errors in observed outcomes or covariates. We also considered the special case of randomized trials.

If treatment assignment is randomized and independent of data errors, ANCOVA estimates of the treatment effect will be unbiased and generally less variable than the moment or multiple imputation estimates described here that incorporate the audit data. One would expect treatment assignment to be independent of data errors in a double-blinded RCT. In this setting, the best approach to estimate the treatment effect is to use the audit data to update the database by correcting any discovered data errors but then to perform a standard analysis on the updated data.

If treatment assignment is correlated with error-prone covariates and/or the covariate errors, then ANCOVA estimates of the treatment effect will be biased, with the extent of the bias depending on the rate and magnitude of the errors, as well as the strength of the correlation between treatment assignment and the errors or covariates. In a double-blinded randomized trial, it is unlikely that treatment assignment would be correlated with data errors. However, such correlation is conceivable in the context of some unblinded randomized trials. For example, consider a trial of two or more treatments that require different specialists. If data-recording errors differ by specialty, then such a correlation may be induced. In addition, many phase II clinical trials are not randomized, so bias due to data errors in these settings could be of concern.

In all of the study settings considered here, standard estimates of an error-prone covariate effect will be biased. Therefore, if estimation of the covariate effect is important, methods to incorporate audit results into the analysis should be considered. In our simulations, we saw that both the moment estimators and our new multiple imputation estimators reduced bias. However, we also saw highly variable estimates with small audit sizes. Large audits may be necessary to get reasonably precise estimates if there are substantial data errors. These findings favor two-stage auditing, where the first audit is used to check for errors, and the second audit is performed, if necessary, to gain more information on the rate and magnitude of errors. Such an approach was discussed and applied by Shepherd and Yu.

Our multiple imputation estimators performed particularly well. In traditional measurement error problems, others have also seen that multiple imputation approaches compare favorably to other methods [12]; however, multiple imputation has been shown to be sensitive to model misspecification [13]. Our data audit setting can be framed as a

measurement error problem, although it differs from most applications because only a portion of the records have data errors and errors are possible, and potentially correlated, in both the outcome and predictor variables. Further study of multiple imputation and other approaches in the data audit setting is warranted.

We have only studied bias and correction methods in settings with a continuous outcome. In clinical trials, binary or time-to-event outcomes are particularly common. Although we suspect that many of the same principles apply, approaches for incorporating audit results in these other settings warrant further research.

Funding

This study was supported in part by the National Institutes of Health (grant numbers 1R01 AI093234-01 and 2U01 AI069923-06).

Conflict of interest

None declared.

References

1. **Shepherd B, Yu C.** Accounting for data errors discovered from an audit in multiple linear regression. *Biometrics* 2011; **67**: 1083–91.
2. **Fuller W.** *Measurement Error Models*. John Wiley & Sons, New York, 1987.
3. **Carroll R, Ruppert D, Stefanski L, Crainiceanu C.** *Measurement Error in Nonlinear Models*. Chapman & Hall, Boca Raton, FL, 2006.
4. **Duda S, Shepherd B, Gadd C, Masys D, McGowan C.** Measuring the quality of observational data in an international HIV research network. *PLoS One*, in press.
5. **Senn S.** Covariate imbalance and random allocation in clinical trials. *Stat Med* 1989; **8**(4): 467–75.
6. **Tsiatis A, Davidian M, Zhang M, Lu X.** Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet? Exible approach. *Stat Med* 2008; **27**: 4658–77.
7. **Assmann S, Pocock S, Enos L, Kasten L.** Subgroup analysis and other (mis) uses of baseline data in clinical trials. *Lancet* 2000; **355**(9209): 1064–69.
8. **Pocock S, Assmann S, Enos L, Kasten L.** Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med* 2002; **21**(19): 2917–30.
9. **Little R, Rubin D.** *Statistical Analysis with Missing Data*. Wiley, New York, 2002.
10. **Carroll R.** Covariance analysis in generalized linear measurement error models. *Stat Med* 1989; **8**: 1075–93.
11. **Carroll R, Gallo P, Gleser L.** Comparison of least squares and errors-in-variables regression, with special reference to randomized analysis of covariance. *J Am Stat Assoc* 1985; **80**: 929–32.
12. **Cole S, Chu H, Greenland S.** Multiple-imputation for measurement-error correction. *Int J Epidemiol* 2006; **35**: 1074–81.
13. **Carpenter J, Kenward M, Vansteelandt S.** A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *J R Stat Soc Ser A* 2006; **169**: 571–84.

Appendix

A multiple imputation approach for incorporating audit findings

In this section, we outline a multiple imputation approach for estimating (β_1,β_2). First, for those with $V = 1$, fit a model of X on W, Z, and Y^*. Using this model, compute the fitted distribution of X for those with $V = 0$. Randomly draw X^{imp} from this fitted distribution. Next, for those with $V = 1$, fit a model of Y on X, W, Z, and Y^*. Using this model, compute the fitted distribution of Y for those with $V = 0$. Then, randomly draw Y^{imp} from this fitted distribution. For those with $V = 0$, the complete imputed data is $(X^{imp},Y^{imp},W,Z,Y^*,V)$, and this is combined with the complete data for those with $V = 1$ to create a complete dataset $(X^{comp},Y^{comp},W,Z,Y^*,V)$, which includes all records. This complete dataset is then analyzed. Specifically, fit the model

$$E\left(Y^{comp}\middle|X^{comp},Z\right)=\alpha_0+\alpha_1X^{comp}+\alpha_2Z$$

and record the estimated slopes, $\hat{\alpha}=(\hat{\alpha}_1,\hat{\alpha}_2)$ and their variance, $\hat{q}=(\hat{q}_1,\hat{q}_2)$. This process is then repeated M times, resulting in M estimates labeled $\hat{\alpha}_i=(\hat{\alpha}_{1i},\hat{\alpha}_{2i})$ and $\hat{q}_i=(\hat{q}_{1i},\hat{q}_{2i})$ for $i = 1,\ldots,M$. The multiple imputation estimators of (β_1,β_2) are then

$$\hat{\beta}_j=\frac{1}{M}\sum_{i=1}^{M}\hat{\alpha}_{ji}$$

for $j = 1,2$, and their variance is estimated as

$$\widehat{var}\left(\hat{\beta}_j\right)=\frac{1}{M}\sum_{i=1}^{M}\hat{q}_{ji}+\frac{M+1}{M}\sum_{i=1}^{M}\left(\hat{q}_{ji}-\frac{1}{M}\sum_{k=1}^{M}\hat{q}_{jk}\right)^2.$$

This multiple imputation procedure can be easily altered to take advantage of different model assumptions. There is some flexibility for fitting models of (X,Y) for those with $V = 0$. The most simple approach is probably to ignore the fact that a large proportion of the records have no errors (i.e., $X = W$ and $Y = Y^*$). The residuals from this model are then a point mass at 0 and some distribution of residuals for those records with errors. One can then impute values by sampling error from either the residuals or from some distribution centered at zero with variance equal to the residual variance.

Table : (Supplementary Material) Performance of estimators under different scenarios relevant for different types of clinical studies, including a multiple imputation approach that samples observed residuals

Estimates of β_2

			Moment-based			MI with normality			MI with random residual		
Scenario[a]	N	n_v	%-Bias	Coverage	MSE$\times 10^3$	%-Bias	Coverage	MSE$\times 10^3$	%-Bias	Coverage	MSE$\times 10^3$
A.	1000	25	-0.66	0.722	52.6	-0.49	0.941	18.7	-0.48	0.943	18.9
		50	-0.45	0.752	19.1	-0.44	0.949	9	-0.42	0.947	9.1
		100	0	0.785	8.5	-0.04	0.946	4.9	-0.03	0.95	4.9
		300	-0.13	0.88	3.1	-0.09	0.949	2.1	-0.09	0.945	2.1
		500	-0.15	0.926	1.9	-0.06	0.951	1.4	-0.05	0.952	1.4
B.	1000	25	-1.53	0.65	69	0.46	0.917	25.9	0.45	0.92	26.2
		50	-0.58	0.695	26.3	-0.16	0.934	11.5	-0.21	0.943	11.5
		100	0.21	0.751	12	0.2	0.93	6.3	0.19	0.932	6.3
		300	-0.11	0.869	3.7	-0.03	0.947	2.3	-0.02	0.952	2.3
		500	-0.03	0.938	2.1	-0.06	0.951	1.6	-0.07	0.951	1.6
C.	1000	25	-8.69	0.734	770.5	2.6	0.919	34.5	1.76	0.918	34.7
		50	-2.94	0.816	67	0.5	0.92	17	0.12	0.923	17
		100	-1.26	0.858	25.2	0.23	0.938	7.6	0.03	0.94	7.6
		300	-0.29	0.919	6.9	-0.1	0.938	2.9	-0.14	0.949	2.9
		500	-0.08	0.935	4	-0.06	0.945	2	-0.08	0.949	2

Estimates of β_1

			Moment-based			MI with normality			MI with random residual		
Scenario[a]	N	n_v	%-Bias	Coverage	MSE$\times 10^7$	%-Bias	Coverage	MSE$\times 10^7$	%-Bias	Coverage	MSE$\times 10^7$
A.	1000	25	7.13	0.831	183	-3.27	0.894	27.2	-1.32	0.889	28
		50	2.86	0.887	58.8	-1.71	0.914	12.6	-0.7	0.912	12.7
		100	1.09	0.915	24.5	-0.54	0.897	7.4	-0.04	0.891	7.5
		300	0.23	0.947	6.3	-0.31	0.927	2.4	-0.19	0.923	2.5
		500	0.01	0.952	3.8	-0.28	0.93	1.7	-0.23	0.929	1.7
B.	1000	25	9.43	0.887	317.1	-2.02	0.906	33.3	-0.25	0.894	34.3
		50	2.98	0.92	69.7	-0.87	0.9	15.2	-0.01	0.894	15.6
		100	1.5	0.944	24.4	-0.07	0.922	7.6	0.36	0.916	7.7
		300	0.45	0.96	7.6	-0.07	0.923	2.6	0.06	0.928	2.7
		500	0.42	0.953	4.3	0.09	0.955	1.6	0.15	0.956	1.6
C.	1000	25	14.58	0.891	1496.6	-1.84	0.905	33.7	-0.17	0.89	35.1
		50	5	0.932	104.6	-0.43	0.913	16	0.36	0.897	16.3
		100	1.41	0.948	34.3	-0.43	0.921	7.4	-0.06	0.913	7.5
		300	0.18	0.949	9.6	-0.2	0.938	2.5	-0.11	0.936	2.5
		500	0.03	0.938	5.3	-0.11	0.945	1.6	-0.07	0.947	1.7

[a] A. Double-blinded randomized trial. B. Unblinded randomized trial. C. Non-randomized study. Details in text.